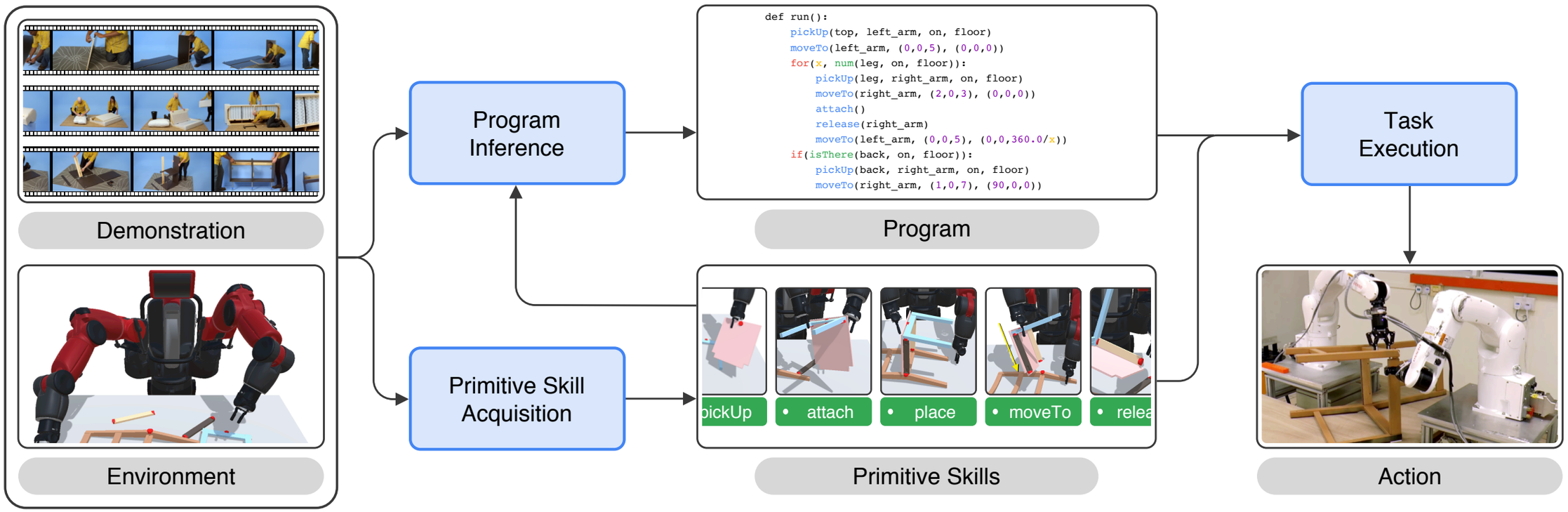
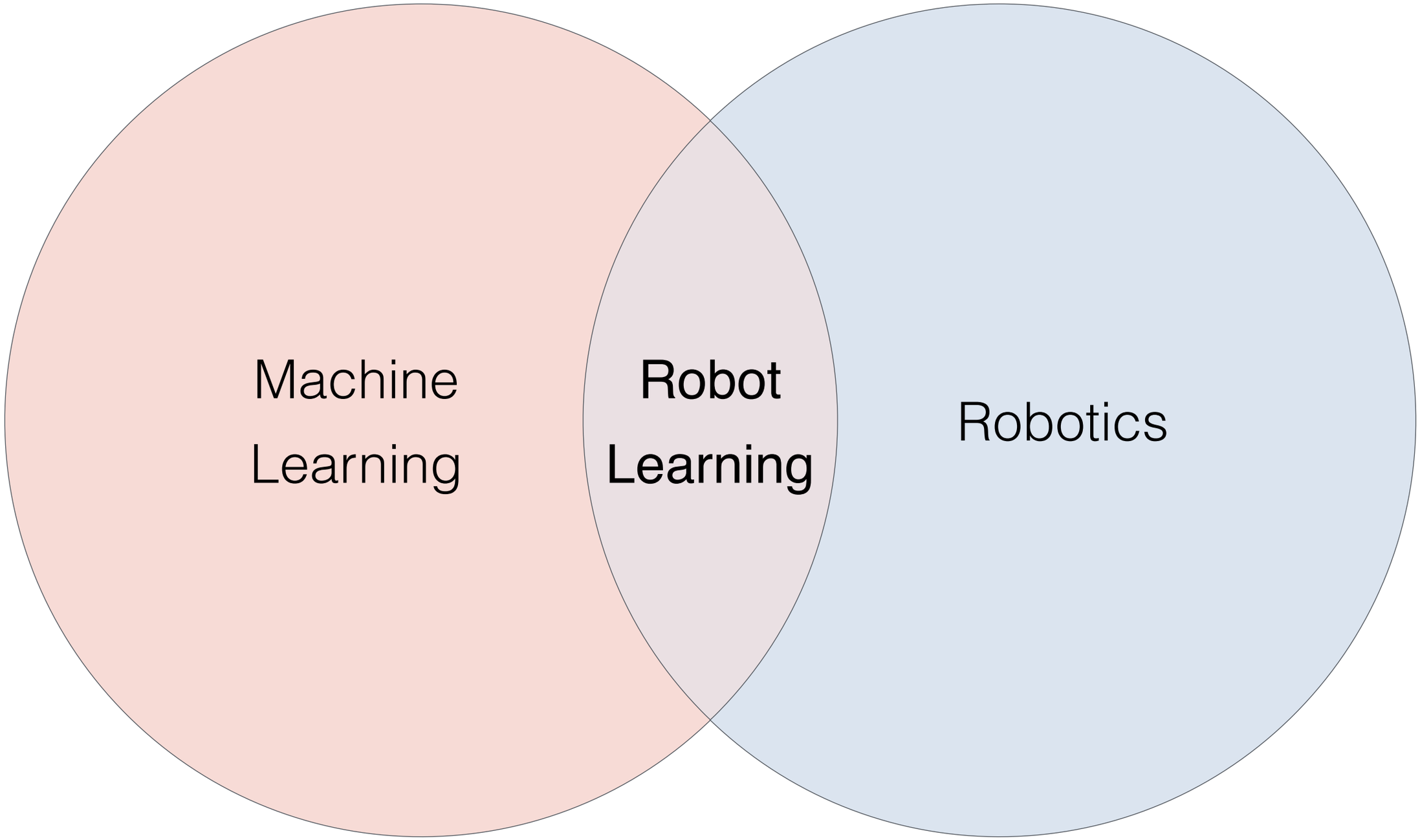
Program-Guided Framework for Interpreting and Acquiring Complex Skills with Learning Robots

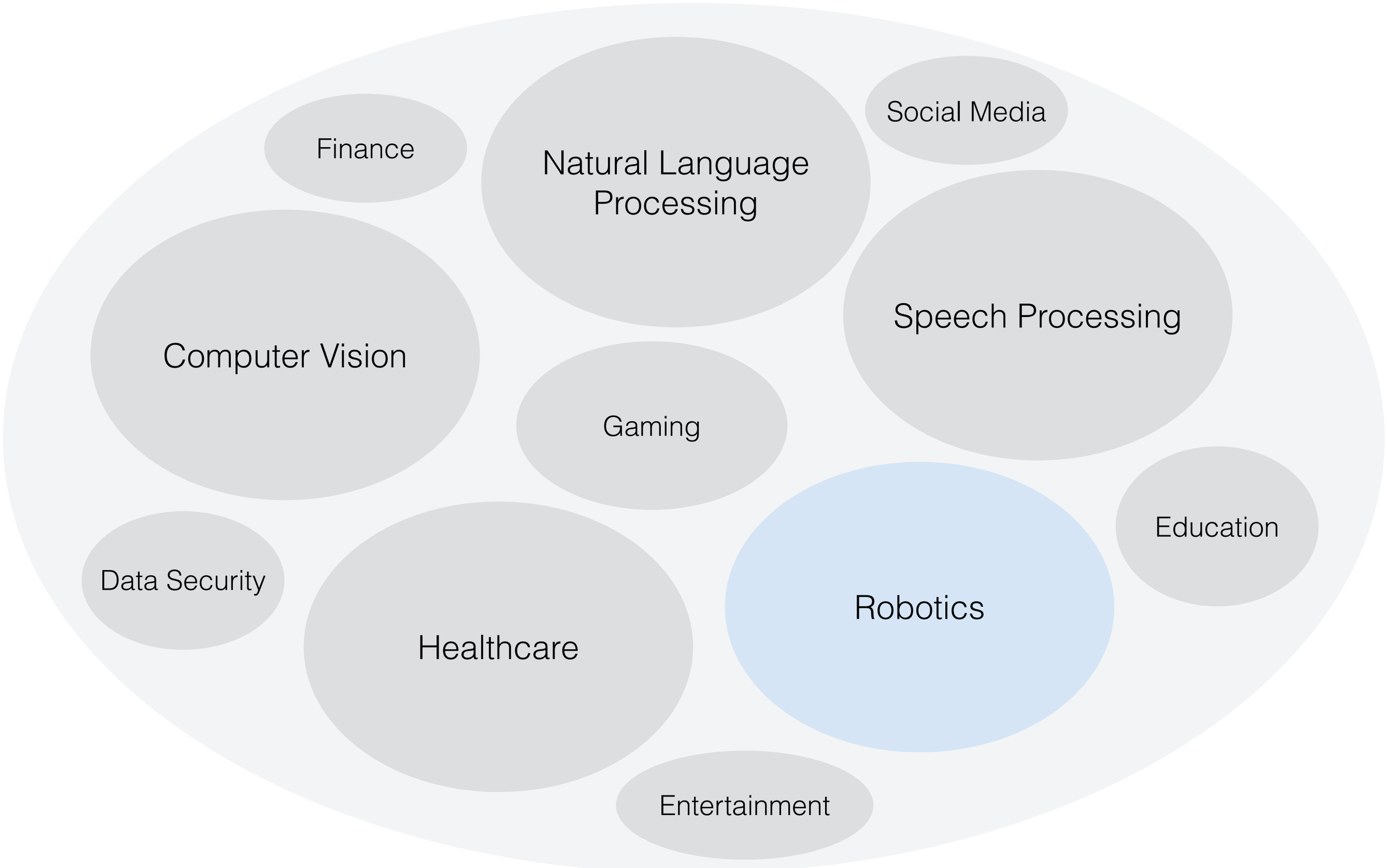


Shao-Hua Sun (孫紹華)

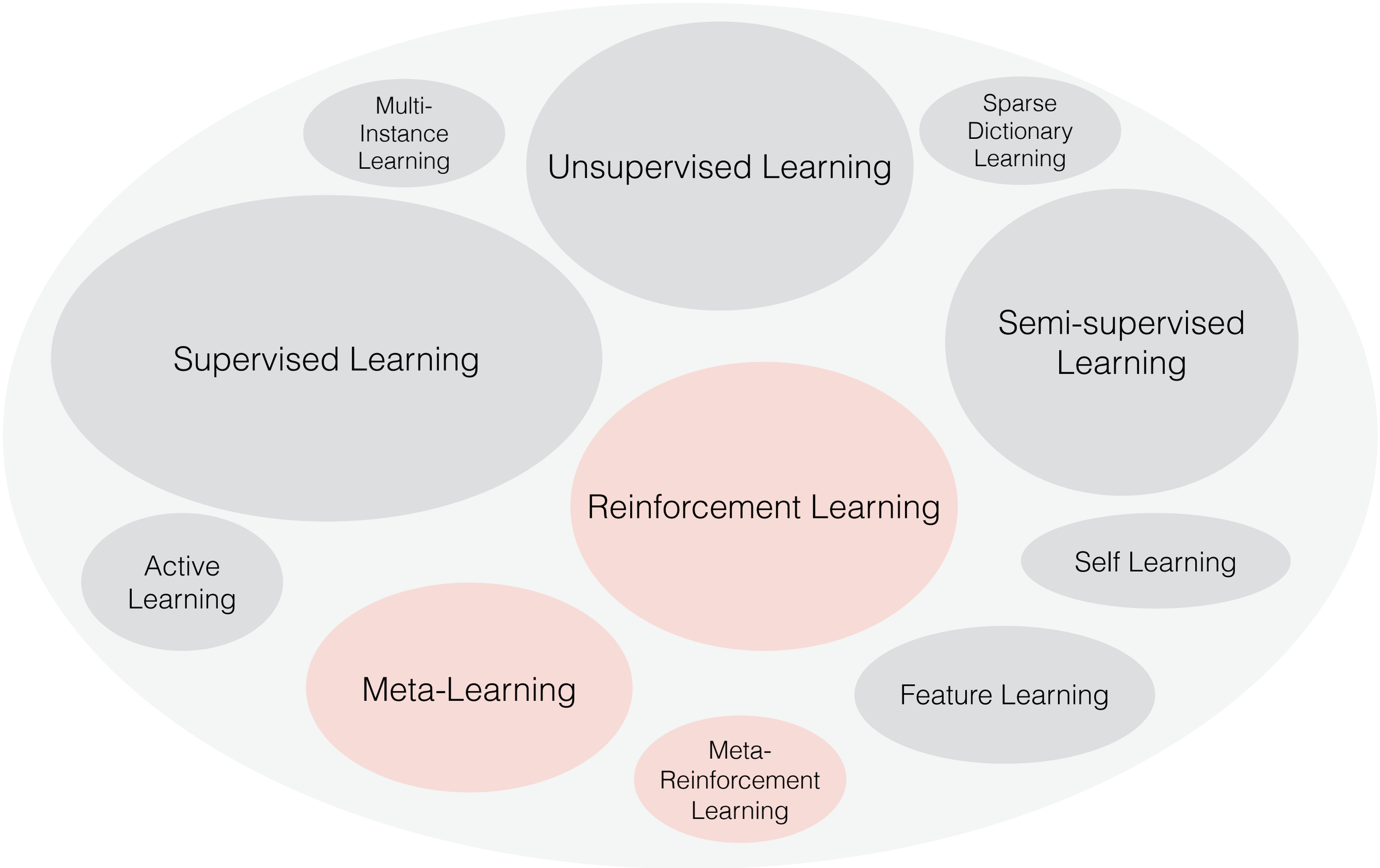
Ph.D. candidate in Computer Science
at the University of Southern California (USC)


Machine
Learning
Robot
Learning
Robotics

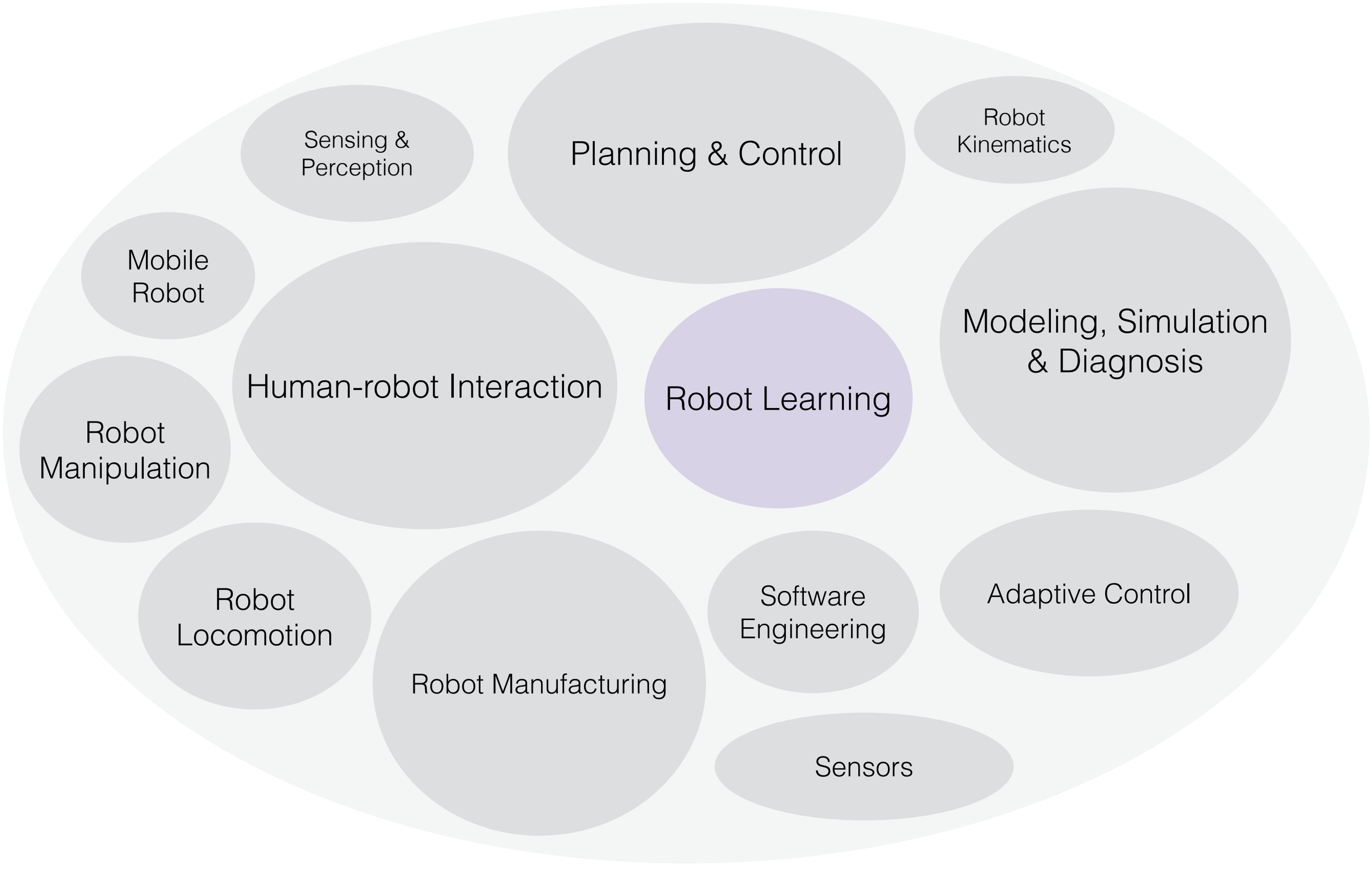
Applications of AI

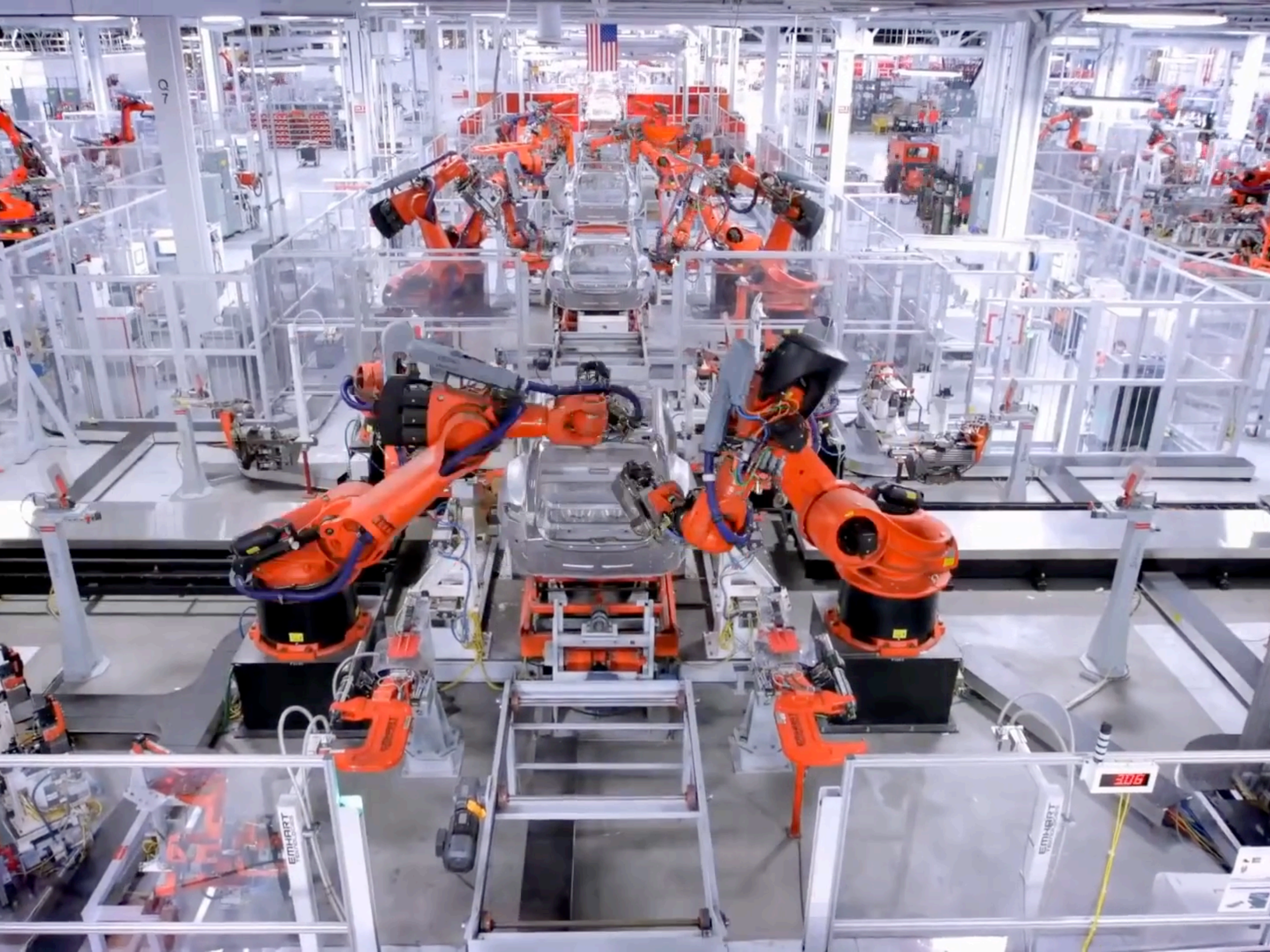


Learning Problems



Robotics

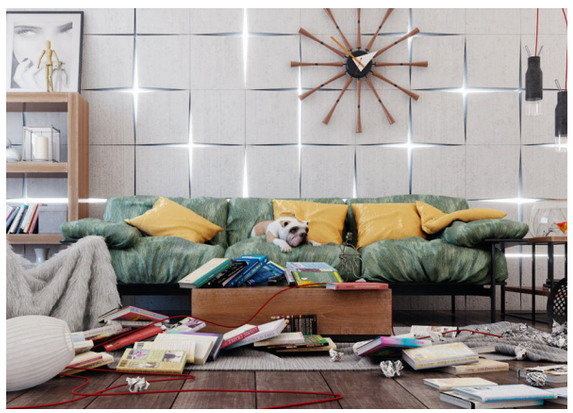

Sensing & Perception
Planning & Control
Robot Kinematics
Mobile Robot
Modeling, Simulation & Diagnosis
Human-robot Interaction
Robot Learning
Robot Manipulation
Robot Locomotion
Robot Manufacturing
Software Engineering
Adaptive Control
Sensors


C7
EMHART
306

Robot Learning

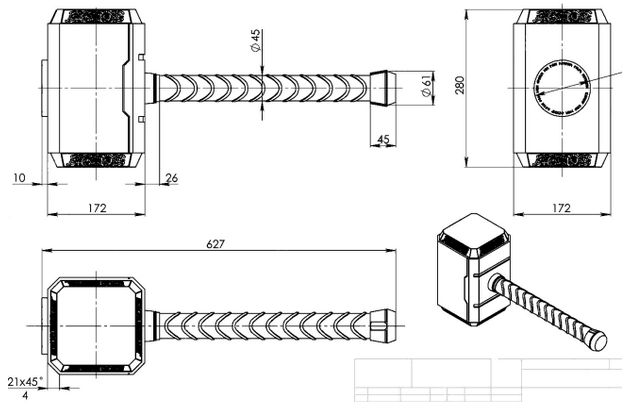
Environment

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 x 4d) [68]	320×320	83.6 M	31.5 B	80.9	95.6
PolyNet [69]	331×331	92 M	34.7 B	81.3	95.8
DPN-131 [8]	320×320	79.5 M	32.0 B	81.5	95.8
SENet [25]	**320×320**	**145.8 M**	**42.3 B**	**82.7**	**96.2**
NASNet-A (6 @ 4032)	**331×331**	**88.9 M**	**23.8 B**	**82.7**	**96.2**

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

Zoph *et al.* Learning Transferable Architectures for Scalable Image Recognition

Instance Segmentation

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

He *et al.* Mask R-CNN

Visual Question Answering

Method	VQA v2 test-dev All	Yes/no	Numb.	Other	VQA v2 test-std All	Yes/no	Numb.	Other
Prior (most common answer in training set) [14]	–	–	–	–	25.98	61.20	0.36	1.17
LSTM Language only (blind model) [14]	–	–	–	–	44.26	67.01	31.55	27.37
Deeper LSTM Q norm. I [?] as reported in [14]	–	–	–	–	54.22	73.46	35.18	41.83
MCB [13] as reported in [14]	–	–	–	–	62.27	78.82	38.28	53.36
UPMC-LIP6 [?]	–	–	–	–	65.71	82.07	41.06	57.12
Athena	–	–	–	–	67.59	82.50	44.19	59.97
LV-NUS	–	–	–	–	66.77	81.89	46.29	58.30
HDU-USYD-UNCC	–	–	–	–	68.09	84.50	45.39	59.01
Proposed model								
ResNet features 7×7, single network	62.07	79.20	39.46	52.62	62.27	79.32	39.77	52.59
Image features from bottom-up attention, adaptive K, single network	65.32	81.82	44.21	56.05	65.67	82.20	43.90	56.26
ResNet features 7×7, ensemble	66.34	83.38	43.17	57.10	66.73	83.71	43.77	57.20
Image features from bottom-up attention, adaptive K, ensemble	**69.87**	**86.08**	**48.99**	**60.80**	**70.34**	**86.60**	**48.64**	**61.15**

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

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

Machine Translation

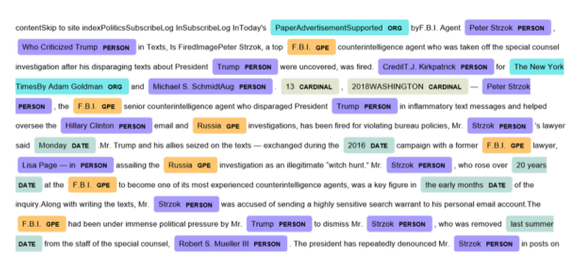
Source	'The reason Boeing are doing this is to cram more seats in to make their plane more competitive with our products,' said Kevin Keniston, head of passenger comfort at Europe's Airbus.
PBMT	'La raison pour laquelle Boeing sont en train de faire, c'est de concentrer davantage de sièges pour prendre leur avion plus compétitive avec nos produits', a déclaré Kevin M. Keniston, chef du confort des passagers de l'Airbus de l'Europe. 3.0
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. 6.0
Human	'Boeing fait ça pour pouvoir caser plus de sièges et rendre ses avions plus compétitifs par rapports à nos produits', a déclaré Kevin Keniston, directeur de Confort Passager chez l'avionneur européen Airbus. 6.0
Source	When asked about this, an official of the American administration replied: 'The United States is not conducting electronic surveillance aimed at offices of the World Bank and IMF in Washington.'
PBMT	Interrogé à ce sujet, un responsable de l'administration américaine a répondu : 'Les Etats-Unis n'est pas effectuer une surveillance électronique destiné aux bureaux de la Banque mondiale et du FMI à Washington'. 3.0
GNMT	Interrogé à ce sujet, un fonctionnaire de l'administration américaine a répondu: 'Les États-Unis n'effectuent pas de surveillance électronique à l'intention des bureaux de la Banque mondiale et du FMI à Washington'. 6.0
Human	Interrogé sur le sujet, un responsable de l'administration américaine a répondu: 'les Etats-Unis ne mènent pas de surveillance électronique visant les sièges de la Banque mondiale et du FMI à Washington'. 6.0

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

Word Embeddings

On word embeddings - Part 1 by Ruder

Named Entity Recognition



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

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	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
#1 Single - nlnet	-	-	83.5	90.1
#2 Single - QANet	-	-	82.5	89.3
Published				
BiDAF+ELMo (Single)	-	85.8	-	-
R.M. Reader (Single)	78.9	86.3	79.5	86.6
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours				
$\text{BERT}_{\text{BASE}}$ (Single)	80.8	88.5	-	-
$\text{BERT}_{\text{LARGE}}$ (Single)	84.1	90.9	-	-
$\text{BERT}_{\text{LARGE}}$ (Ensemble)	85.8	91.8	-	-
$\text{BERT}_{\text{LARGE}}$ (Sgl.+TriviaQA)	**84.2**	**91.1**	**85.1**	**91.8**
$\text{BERT}_{\text{LARGE}}$ (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

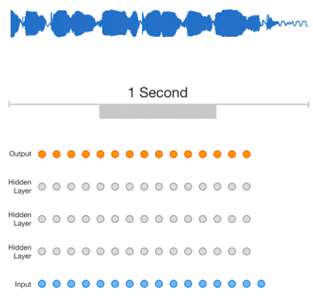
Speech Recognition

Word Error Rate

Senone set	Model/combination step	WER devset ngram-LM	WER test ngram-LM	WER devset LSTM-LMs	WER test LSTM-LMs
9k	BLSTM	11.5	8.3	9.2	6.3
27k	BLSTM	11.4	8.0	9.3	6.3
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]	224×224	11.2 M	1.94 B	74.8	92.2
NASNet-A (5 @ 1538)	**299×299**	**10.9 M**	**2.35 B**	**78.6**	**94.2**
Inception V3 [60]	299×299	23.8 M	5.72 B	78.8	94.4
Xception [9]	299×299	22.8 M	8.38 B	79.0	94.5
Inception ResNet V2 [58]	299×299	55.8 M	13.2 B	80.1	95.1
NASNet-A (7 @ 1920)	**299×299**	**22.6 M**	**4.93 B**	**80.8**	**95.3**
ResNeXt-101 (64 x 4d) [68]	320×320	83.6 M	31.5 B	80.9	95.6
PolyNet [69]	331×331	92 M	34.7 B	81.3	95.8
DPN-131 [8]	320×320	79.5 M	32.0 B	81.5	95.8
SENet [25]	**320×320**	**145.8 M**	**42.3 B**	**82.7**	**96.2**
NASNet-A (6 @ 4032)	**331×331**	**88.9 M**	**23.8 B**	**82.7**	**96.2**

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

Zoph *et al.* Learning Transferable Architectures for Scalable Image Recognit

Instance Segmentation

Fig. 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

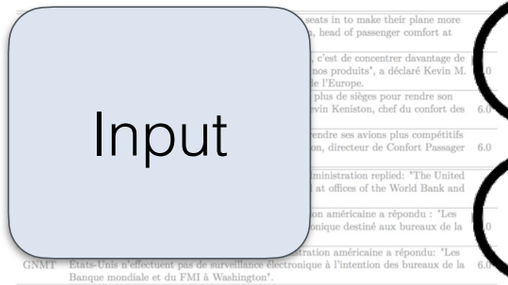
Visual Question Answering

Method	VQA v2 test-dev All	Yes/no	Num.	Other	VQA v2 test-std All	Yes/no	Num.	Other
Prior (most common answer in training set) [14]	–	–	–	–	25.98	61.20	0.36	1.17
LSTM Language only (blind model) [14]	–	–	–	–	44.26	67.01	31.55	27.37
Deeper LSTM Q norm. I [7] as reported in [14]	–	–	–	–	54.22	73.46	35.18	41.83
MCB [7] as reported in [14]	–	–	–	–	62.27	78.82	38.28	53.36
UPMC-LIP6 [7]	–	–	–	–	65.71	82.07	41.06	57.12
Athena	–	–	–	–	67.59	82.50	44.19	59.97
LV-NUS	–	–	–	–	66.77	81.89	46.29	58.30
HDU-USYD-UNCC	–	–	–	–	68.09	84.50	45.39	59.01
Proposed model								
ResNet features 7×7, single network	62.07	79.20	39.46	52.62	62.27	79.32	39.77	52.59
Image features from bottom-up attention, adaptive K, single network	65.32	81.82	44.21	56.05	65.67	82.20	43.90	56.26
ResNet features 7×7, ensemble	66.34	83.38	43.17	57.10	66.73	83.71	43.77	57.20
Image features from bottom-up attention, adaptive K, ensemble	**69.87**	**86.08**	**48.99**	**60.80**	**70.34**	**86.60**	**48.64**	**61.15**

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

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

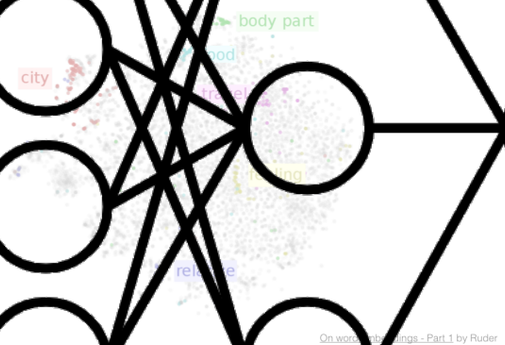
Machine Translation



GNMT	Etats-Unis n'effectuent pas de surveillance électronique à l'intention des bureaux de la Banque mondiale et du FMI à Washington".
Human	Interrogé sur le sujet, un responsable de l'administration américaine a répondu: "les Etats-Unis ne mènent pas de surveillance électronique visant les sièges de la Banque mondiale et du FMI à Washington".

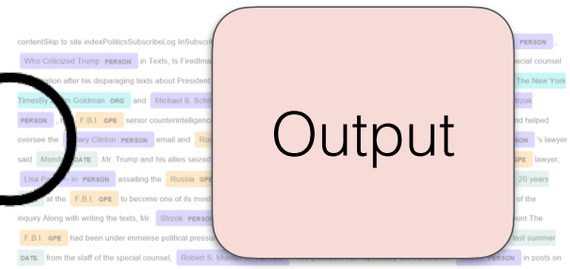
Wu *et al.* Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation

Word Embeddings



On word embeddings - Part 1 by Ruder

Named Entity Recognition



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

- 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	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
#1 Single - nlnet	-	-	83.5	90
#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.+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

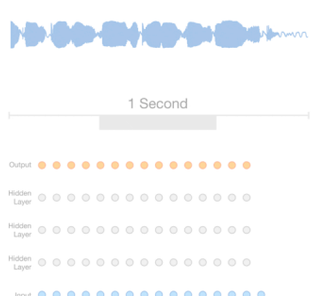
Speech Recognition

Word Error Rate

Senone set	Model/combination step	WER devset ngram-LM	WER test ngram-LM	WER devset LSTM-LMs	WER test LSTM-LMs
9k	BLSTM	11.5	8.3	9.2	6.3
27k	BLSTM	11.4	8.0	9.3	6.3
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

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 x 4d) [68]	320×320	83.6 M	31.5 B	80.9	95.6
PolyNet [69]	331×331	92 M	34.7 B	81.3	95.8
DPN-131 [8]	320×320	79.5 M	32.0 B	81.5	95.8
SENet [25]	**320×320**	**145.8 M**	**42.3 B**	**82.7**	**96.2**
NASNet-A (6 @ 4032)	**331×331**	**88.9 M**	**23.8 B**	**82.7**	**96.2**

Table 2. Performance of architecture search and other published state-of-the-art models on ImageNet classification. Mult-Adds indicate the number of composite multiply-accumulate operations for a single image. Note that the composite multiple-accumulate operations are calculated for the image size reported in the table. Models in [6, 25] calculated from open-source implementation.

Zoph *et al.* Learning Transferable Architectures for Scalable Image Recognit

Instance Segmentation

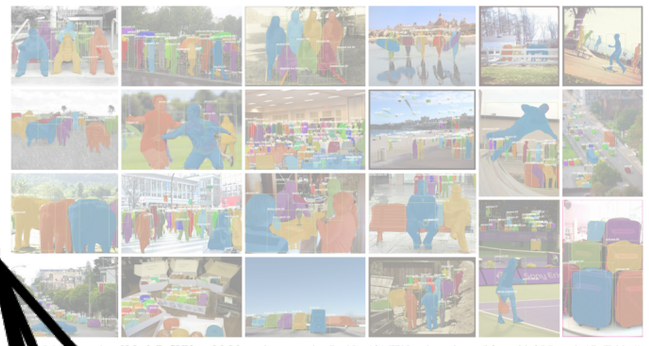


Fig. 5. Mask results of **Mask R-CNN** on COCO test images, using ResNet-101-FPN and running at 5 fps, with 35.7 mask AP (Table 1).

He *et al.* Mask R-CNN

Visual Question Answering

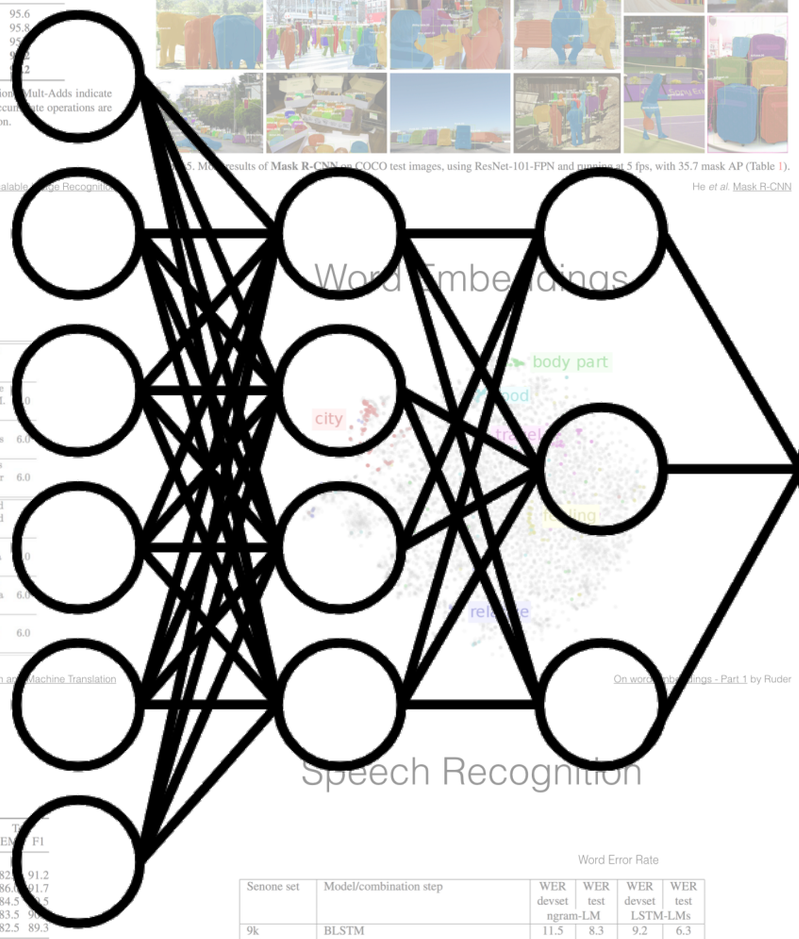
Method	VQA v2 test-dev All	Yes/no	Num.	Other	VQA v2 test-std All	Yes/no	Num.	Other
Prior (most common answer in training set) [14]	–	–	–	–	25.98	61.20	0.36	1.17
LSTM Language only (blind model) [14]	–	–	–	–	44.26	67.01	31.55	27.37
Deeper LSTM Q norm. I [7] as reported in [14]	–	–	–	–	54.22	73.46	35.18	41.83
MCB [7] as reported in [14]	–	–	–	–	62.27	78.82	38.28	53.36
UPMC-LIP6 [7]	–	–	–	–	65.71	82.07	41.06	57.12
Athena	–	–	–	–	67.59	82.50	44.19	59.97
LV-NUS	–	–	–	–	66.77	81.89	46.29	58.30
HDU-USYD-UNCC	–	–	–	–	68.09	84.50	45.39	59.01
Proposed model								
ResNet features 7×7, single network	62.07	79.20	39.46	52.62	62.27	79.32	39.77	52.59
Image features from bottom-up attention, adaptive K, single network	65.32	81.82	44.21	56.05	65.67	82.20	43.90	56.26
ResNet features 7×7, ensemble	66.34	83.38	43.17	57.10	66.73	83.71	43.77	57.20
Image features from bottom-up attention, adaptive K, ensemble	**69.87**	**86.08**	**48.99**	**60.80**	**70.34**	**86.60**	**48.64**	**61.15**

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

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

Image

Word Embeddings



On word embeddings - Part 1 by Ruder

Named Entity Recognition

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

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	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
#1 Single - nlnet	-	-	83.5	90.1
#2 Single - QANet	-	-	82.5	89.3
Published				
BiDAF+ELMo (Single)	-	85.8	-	-
R.M. Reader (Single)	78.9	86.3	79.5	86.6
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours				
$\text{BERT}_{\text{BASE}}$ (Single)	80.8	88.5	-	-
$\text{BERT}_{\text{LARGE}}$ (Single)	84.1	90.9	-	-
$\text{BERT}_{\text{LARGE}}$ (Ensemble)	85.8	91.8	-	-
$\text{BERT}_{\text{LARGE}}$ (Sgl.+TriviaQA)	**84.2**	**91.1**	**85.1**	**91.8**
$\text{BERT}_{\text{LARGE}}$ (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

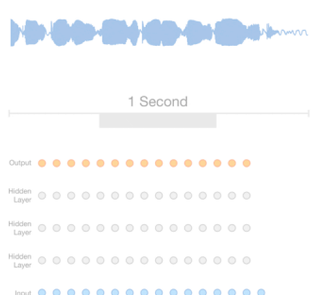
Speech Recognition

Word Error Rate

Senone set	Model/combination step	WER devset ngram-LM	WER test ngram-LM	WER devset LSTM-LMs	WER test LSTM-LMs
9k	BLSTM	11.5	8.3	9.2	6.3
27k	BLSTM	11.4	8.0	9.3	6.3
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

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 x 4d) [68]	320×320	83.6 M	31.5 B	80.9	95.6
PolyNet [69]	331×331	92 M	34.7 B	81.3	95.8
DPN-131 [8]	320×320	79.5 M	32.0 B	81.5	95.8
SENet [25]	**320×320**	**145.8 M**	**42.3 B**	**82.7**	**96.2**
NASNet-A (6 @ 4032)	**331×331**	**88.9 M**	**23.8 B**	**82.7**	**96.2**

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

Zoph *et al.* Learning Transferable Architectures for Scalable Image Recognit

Instance Segmentation

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

He *et al.* Mask R-CNN

Visual Question Answering

Method	VQA v2 test-dev All	Yes/no	Num.	Other	VQA v2 test-std All	Yes/no	Num.	Other
Prior (most common answer in training set) [14]	–	–	–	–	25.98	61.20	0.36	1.17
LSTM Language only (blind model) [14]	–	–	–	–	44.26	67.01	31.55	27.37
Deeper LSTM Q norm. I [7] as reported in [14]	–	–	–	–	54.22	73.46	35.18	41.83
MCB [7] as reported in [14]	–	–	–	–	62.27	78.82	38.28	53.36
UPMC-LIP6 [7]	–	–	–	–	65.71	82.07	41.06	57.12
Athena	–	–	–	–	67.59	82.50	44.19	59.97
LV-NUS	–	–	–	–	66.77	81.89	46.29	58.30
HDU-USYD-UNCC	–	–	–	–	68.09	84.50	45.39	59.01
Proposed model								
ResNet features 7×7, single network	62.07	79.20	39.46	52.62	62.27	79.32	39.77	52.59
Image features from bottom-up attention, adaptive K, single network	65.32	81.82	44.21	56.05	65.67	82.20	43.90	56.26
ResNet features 7×7, ensemble	66.34	83.38	43.17	57.10	66.73	83.71	43.77	57.20
Image features from bottom-up attention, adaptive K, ensemble	**69.87**	**86.08**	**48.99**	**60.80**	**70.34**	**86.60**	**48.64**	**61.15**

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

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

English sentence

French sentence

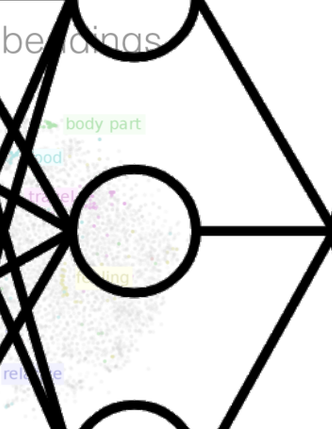
Machine Translation

France is never cold in September

GNMT	Etats-Unis n'effectuent pas de surveillance électronique à l'intention des bureaux de la Banque mondiale et du FMI à Washington".
Human	Interrogé sur le sujet, un responsable de l'administration américaine a répondu: "les Etats-Unis ne mènent pas de surveillance électronique visant les sièges de la Banque mondiale et du FMI à Washington".

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

Word Embeddings



On word embeddings - Part 1 by Ruder

Named Entity Recognition

la france est jamais froid en septembre

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

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

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	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
#1 Single - nlnet	-	-	83.5	90.1
#2 Single - QANet	-	-	82.5	89.3
Published				
BiDAF+ELMo (Single)	-	85.8	-	-
R.M. Reader (Single)	78.9	86.3	79.5	86.6
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours				
$BERT_{BASE}$ (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**
$BERT_{LARGE}$ (Ens.+TriviaQA)	**86.2**	**92.2**	**87.4**	**93.2**

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

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

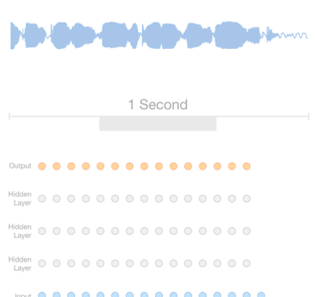
Speech Recognition

Word Error Rate

Senone set	Model/combination step	WER devset ngram-LM	WER test ngram-LM	WER devset LSTM-LMs	WER test LSTM-LMs
9k	BLSTM	11.5	8.3	9.2	6.3
27k	BLSTM	11.4	8.0	9.3	6.3
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

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 x 4d) [68]	320×320	83.6 M	31.5 B	80.9	95.6
PolyNet [69]	331×331	92 M	34.7 B	81.3	95.8
DPN-131 [8]	320×320	79.5 M	32.0 B	81.5	95.8
SENet [25]	**320×320**	**145.8 M**	**42.3 B**	**82.7**	**96.2**
NASNet-A (6 @ 4032)	**331×331**	**88.9 M**	**23.8 B**	**82.7**	**96.2**

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

Zoph *et al.* Learning Transferable Architectures for Scalable Image Recognit

Instance Segmentation

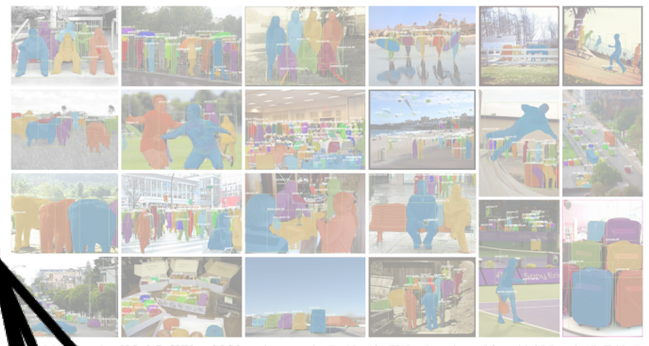


Fig. 5. More results of **Mask R-CNN** on COCO test images, using ResNet-101-FPN and running at 5 fps, with 35.7 mask AP (Table 1).

He *et al.* Mask R-CNN

Visual Question Answering

Method	VQA v2 test-dev All	Yes/no	Num.	Other	VQA v2 test-std All	Yes/no	Num.	Other
Prior (most common answer in training set) [14]	–	–	–	–	25.98	61.20	0.36	1.17
LSTM Language only (blind model) [14]	–	–	–	–	44.26	67.01	31.55	27.37
Deeper LSTM Q norm. I [7] as reported in [14]	–	–	–	–	54.22	73.46	35.18	41.83
MCB [7] as reported in [14]	–	–	–	–	62.27	78.82	38.28	53.36
UPMC-LIP6 [7]	–	–	–	–	65.71	82.07	41.06	57.12
Athena	–	–	–	–	67.59	82.50	44.19	59.97
LV-NUS	–	–	–	–	66.77	81.89	46.29	58.30
HDU-USYD-UNCC	–	–	–	–	68.09	84.50	45.39	59.01
Proposed model								
ResNet features 7×7, single network	62.07	79.20	39.46	52.62	62.27	79.32	39.77	52.59
Image features from bottom-up attention, adaptive K, single network	65.32	81.82	44.21	56.05	65.67	82.20	43.90	56.26
ResNet features 7×7, ensemble	66.34	83.38	43.17	57.10	66.73	83.71	43.77	57.20
Image features from bottom-up attention, adaptive K, ensemble	**69.87**	**86.08**	**48.99**	**60.80**	**70.34**	**86.60**	**48.64**	**61.15**

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

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

Waveform

Machine Translation

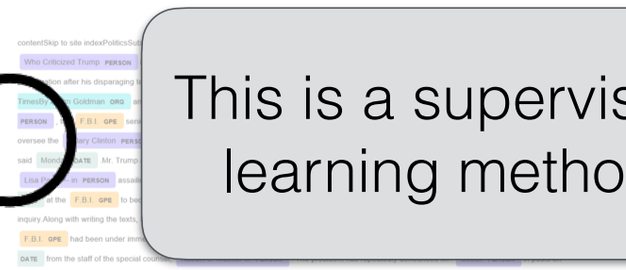
Source	"The reason Boeing are doing this is to cram more seats in to make their plane more competitive with our products," said Kevin Keniston, head of passenger comfort at Europe's Airbus.
PBMT	"La raison pour laquelle [illegible] en train de f[illegible] c'est de concentrer davantage de sièges pour pr[illegible]re leur [illegible] compétitive [illegible] produits", a déclaré Kevin M. Keniston, chef [illegible] confort des passagers chez l'Airbus.
GNMT	"La raison pour laquelle [illegible] pour rendre son [illegible] plus compétitive avec nos produits", a déclaré Kevin Keniston, chef du confort des [illegible] chez l'Airbus.
Source	When asked about this [illegible] American [illegible] replied: "The United States is not conducting [illegible] surveillance [illegible] the World Bank and IMF in Washington."
PBMT	Interrogé à ce sujet, un [illegible] de l'administration américaine a répondu : "Les États-Unis n'est pas effectuant une surveillance électronique destiné aux bureaux de la Banque mondiale et du FMI à Washington".
GNMT	Interrogé à ce sujet, un fonctionnaire de l'administration américaine a répondu: "Les États-Unis n'effectuent pas de surveillance électronique à l'intention des bureaux de la Banque mondiale et du FMI à Washington".
Human	Interrogé sur le sujet, un responsable de l'administration américaine a répondu: "les États-Unis ne mènent pas de surveillance électronique visant les sièges de la Banque mondiale et du FMI à Washington".

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

Word Embeddings

On word embeddings - Part 1 by Ruder

Named Entity Recognition



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

Text

This is a supervised learning method

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	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
#1 Single - nlnet	-	-	83.5	90.1
#2 Single - QANet	-	-	82.5	89.3
Published				
BiDAF+ELMo (Single)	-	85.8	-	-
R.M. Reader (Single)	78.9	86.3	79.5	86.6
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours				
$BERT_{BASE}$ (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**
$BERT_{LARGE}$ (Ens.+TriviaQA)	**86.2**	**92.2**	**87.4**	**93.2**

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

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

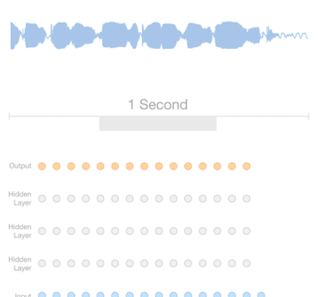
Speech Recognition

Word Error Rate

Senone set	Model/combination step	WER devset ngram-LM	WER test ngram-LM	WER devset LSTM-LMs	WER test LSTM-LMs
9k	BLSTM	11.5	8.3	9.2	6.3
27k	BLSTM	11.4	8.0	9.3	6.3
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]	224×224	11.2 M	1.94 B	74.8	92.2
NASNet-A (5 @ 1538)	**299×299**	**10.9 M**	**2.35 B**	**78.6**	**94.2**
Inception V3 [60]	299×299	23.8 M	5.72 B	78.8	94.4
Xception [9]	299×299	22.8 M	8.38 B	79.0	94.5
Inception ResNet V2 [58]	299×299	55.8 M	13.2 B	80.1	95.1
NASNet-A (7 @ 1920)	**299×299**	**22.6 M**	**4.93 B**	**80.8**	**95.3**
ResNeXt-101 (64 x 4d) [68]	320×320	83.6 M	31.5 B	80.9	95.6
PolyNet [69]	331×331	92 M	34.7 B	81.3	95.8
DPN-131 [8]	320×320	79.5 M	32.0 B	81.5	95.8
SENet [25]	**320×320**	**145.8 M**	**42.3 B**	**82.7**	**96.2**
NASNet-A (6 @ 4032)	**331×331**	**88.9 M**	**23.8 B**	**82.7**	**96.2**

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

Zoph *et al.* Learning Transferable Architectures for Scalable Image Recognit

Instance Segmentation

Fig. 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

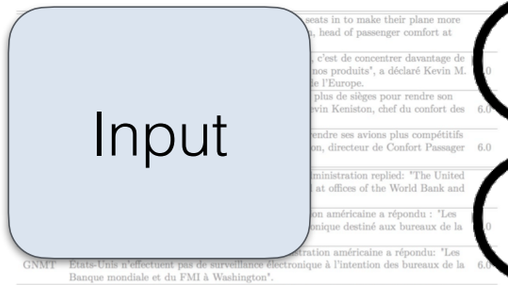
Visual Question Answering

Method	VQA v2 test-dev All	Yes/no	Num.	Other	VQA v2 test-std All	Yes/no	Num.	Other
Prior (most common answer in training set) [14]	–	–	–	–	25.98	61.20	0.36	1.17
LSTM Language only (blind model) [14]	–	–	–	–	44.26	67.01	31.55	27.37
Deeper LSTM Q norm. I [7] as reported in [14]	–	–	–	–	54.22	73.46	35.18	41.83
MCB [7] as reported in [14]	–	–	–	–	62.27	78.82	38.28	53.36
UPMC-LIP6 [7]	–	–	–	–	65.71	82.07	41.06	57.12
Athena	–	–	–	–	67.59	82.50	44.19	59.97
LV-NUS	–	–	–	–	66.77	81.89	46.29	58.30
HDU-USYD-UNCC	–	–	–	–	68.09	84.50	45.39	59.01
Proposed model								
ResNet features 7×7, single network	62.07	79.20	39.46	52.62	62.27	79.32	39.77	52.59
Image features from bottom-up attention, adaptive K, single network	65.32	81.82	44.21	56.05	65.67	82.20	43.90	56.26
ResNet features 7×7, ensemble	66.34	83.38	43.17	57.10	66.73	83.71	43.77	57.20
Image features from bottom-up attention, adaptive K, ensemble	**69.87**	**86.08**	**48.99**	**60.80**	**70.34**	**86.60**	**48.64**	**61.15**

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

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

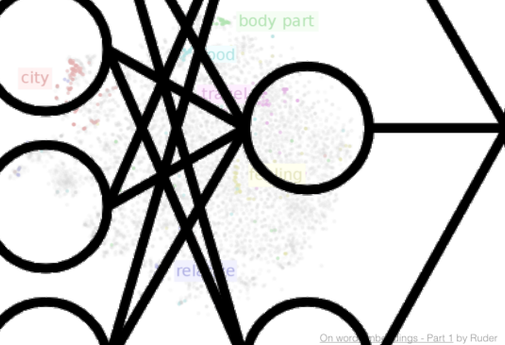
Machine Translation



GNMT	Etats-Unis n'effectuent pas de surveillance électronique à l'intention des bureaux de la Banque mondiale et du FMI à Washington".
Human	Interrogé sur le sujet, un responsable de l'administration américaine a répondu: "les Etats-Unis ne mènent pas de surveillance électronique visant les sièges de la Banque mondiale et du FMI à Washington".

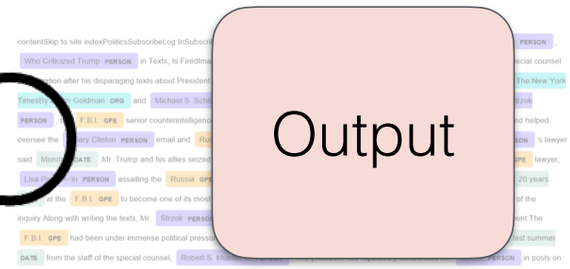
Wu *et al.* Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation

Word Embeddings



On word embeddings - Part 1 by Ruder

Named Entity Recognition



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

- 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	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
#1 Single - nlnet	-	-	83.5	90
#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.+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

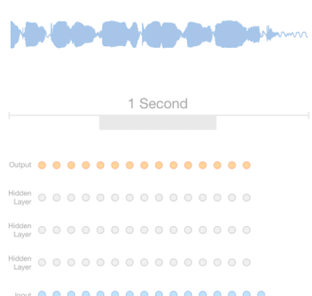
Speech Recognition

Word Error Rate

Senone set	Model/combination step	WER devset ngram-LM	WER test ngram-LM	WER devset LSTM-LMs	WER test LSTM-LMs
9k	BLSTM	11.5	8.3	9.2	6.3
27k	BLSTM	11.4	8.0	9.3	6.3
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]	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 [30]	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	30.4 M	13.2 B	80.1	95.1
NASNet-A (7 @ 1920)	299×299	22.6 M	**4.93 B**	**80.8**	**95.3**
[illegible]	[illegible]	[illegible]	[illegible]	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**	[illegible]
NASNet-A (6 @ 4032)	331×331	**88.9 M**	**23.8 B**	**82.7**	[illegible]

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

Zoph *et al.* Learning Transferable Architectures for Scalable Image Recognition

Instance Segmentation

Fig. 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 All	Yes/no	Num.	Other	VQA v2 test-std All	Yes/no	Num.	Other
Prior (most common answer in training set) [14]	–	–	–	–	25.98	61.20	0.36	1.17
LSTM Language only (blind model) [14]	–	–	–	–	44.26	67.01	31.55	27.37
Deeper LSTM Q norm. I [7] as reported in [14]	–	–	–	–	54.22	73.46	35.18	41.83
MCB [7] as reported in [14]	–	–	–	–	62.27	78.82	38.28	53.36
UPMC-LIP6 [7]	–	–	–	–	65.71	82.07	41.06	57.12
Athena	–	–	–	–	67.59	82.50	44.19	59.97
LV-NUS	–	–	–	–	66.77	81.89	46.29	58.30
HDU-USYD-UNCC	–	–	–	–	68.09	84.50	45.39	59.01
Proposed model								
ResNet features 7×7, single network	62.07	79.20	39.46	52.62	62.27	79.32	39.77	52.59
Image features from bottom-up attention, adaptive K, single network	65.32	81.82	44.21	56.05	65.67	82.20	43.90	56.26
ResNet features 7×7, ensemble	66.34	83.38	43.17	57.10	66.73	83.71	43.77	57.20
Image features from bottom-up attention, adaptive K, ensemble	**69.87**	**86.08**	**48.99**	**60.80**	**70.34**	**86.60**	**48.64**	**61.15**

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

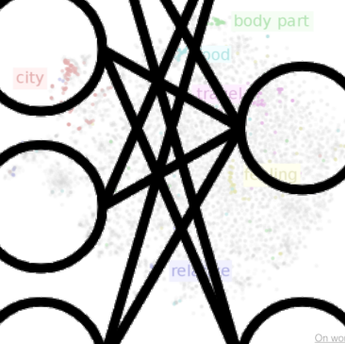
Teney *et al.* Tips and Tricks for Visual Question Answering: Learnings from the 2017 Challenge

Machine Translation

GNMT	Etats-Unis n'effectuent pas de surveillance électronique à l'intention des bureaux de la Banque mondiale et du FMI à Washington".	6.5
Human	Interrogé sur le sujet, un responsable de l'administration américaine a répondu: "les Etats-Unis ne mènent pas de surveillance électronique visant les sièges de la Banque mondiale et du FMI à Washington".	6.0

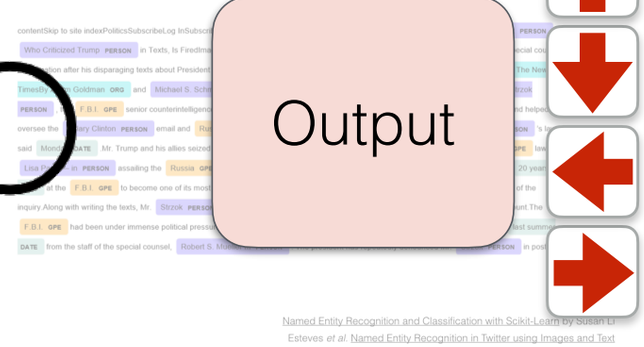
Wu *et al.* Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation

Word Embeddings



On word embeddings - Part 1 by Ruder

Named Entity Recognition



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

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	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
#1 Single - nlnet	-	-	83.5	90.1
#2 Single - QANet	-	-	82.5	89.3
Published				
BiDAF+ELMo (Single)	-	85.8	-	-
R.M. Reader (Single)	78.9	86.3	79.5	86.6
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours				
$BERT_{BASE}$ (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**
$BERT_{LARGE}$ (Ens.+TriviaQA)	**86.2**	**92.2**	**87.4**	**93.2**

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

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

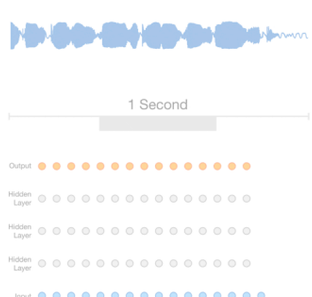
Speech Recognition

Word Error Rate

Senone set	Model/combination step	WER devset ngram-LM	WER test ngram-LM	WER devset LSTM-LMs	WER test LSTM-LMs
9k	BLSTM	11.5	8.3	9.2	6.3
27k	BLSTM	11.4	8.0	9.3	6.3
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]	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 [30]	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	30.4 M	13.2 B	80.1	95.1
NASNet-A (7 @ 1920)	299×299	22.6 M	**4.93 B**	**80.8**	**95.3**
[illegible]	[illegible]	[illegible]	[illegible]	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**	[illegible]
NASNet-A (6 @ 4032)	331×331	**88.9 M**	**23.8 B**	**82.7**	[illegible]

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

Zoph *et al.* Learning Transferable Architectures for Scalable Image Recognition

Instance Segmentation

Fig. 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

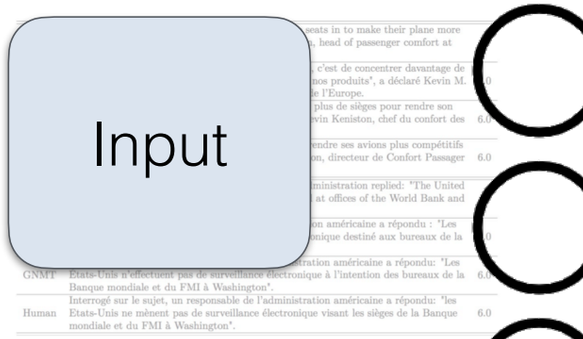
Visual Question Answering

Method	VQA v2 test-dev All	Yes/no	Num.	Other	VQA v2 test-std All	Yes/no	Num.	Other
Prior (most common answer in training set) [14]	–	–	–	–	25.98	61.20	0.36	1.17
LSTM Language only (blind model) [14]	–	–	–	–	44.26	67.01	31.55	27.37
Deeper LSTM Q norm. I [7] as reported in [14]	–	–	–	–	54.22	73.46	35.18	41.83
MCB [7] as reported in [14]	–	–	–	–	62.27	78.82	38.28	53.36
UPMC-LIP6 [7]	–	–	–	–	65.71	82.07	41.06	57.12
Athena	–	–	–	–	67.59	82.50	44.19	59.97
LV-NUS	–	–	–	–	66.77	81.89	46.29	58.30
HDU-USYD-UNCC	–	–	–	–	68.09	84.50	45.39	59.01
Proposed model								
ResNet features 7×7, single network	62.07	79.20	39.46	52.62	62.27	79.32	39.77	52.59
Image features from bottom-up attention, adaptive K, single network	65.32	81.82	44.21	56.05	65.67	82.20	43.90	56.26
ResNet features 7×7, ensemble	66.34	83.38	43.17	57.10	66.73	83.71	43.77	57.20
Image features from bottom-up attention, adaptive K, ensemble	**69.87**	**86.08**	**48.99**	**60.80**	**70.34**	**86.60**	**48.64**	**61.15**

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

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

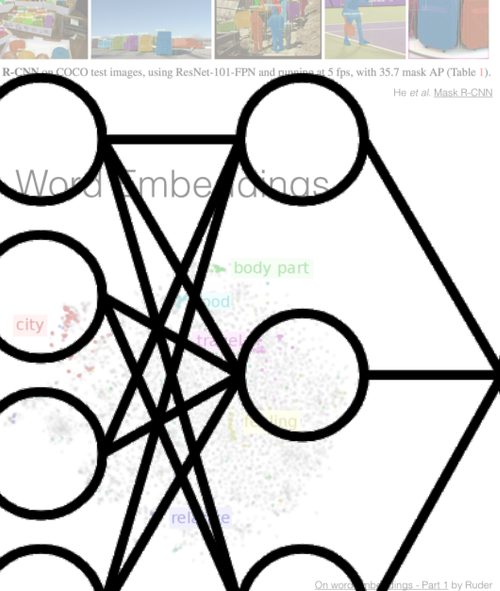
Machine Translation



	[illegible]eats in to make their plane more [illegible], head of passenger comfort at
	[illegible] c'est de concentrer davantage de [illegible]nos produits', a déclaré Kevin M. [illegible]e l'Europe.
	[illegible] plus de sièges pour rendre son [illegible]evin Keniston, chef du confort des
	[illegible]rendre ses avions plus compétitifs [illegible]on, directeur de Confort Passager
	[illegible]ministration replied: "The United [illegible] at offices of the World Bank and
	[illegible]on américaine a répondu : "Les [illegible]nique destiné aux bureaux de la
GNMT	[illegible]tration américaine a répondu: 'Les Etats-Unis n'effectuent pas de surveillance électronique à l'intention des bureaux de la Banque mondiale et du FMI à Washington'.
Human	Interrogé sur le sujet, un responsable de l'administration américaine a répondu: "les Etats-Unis ne mènent pas de surveillance électronique visant les sièges de la Banque mondiale et du FMI à Washington".

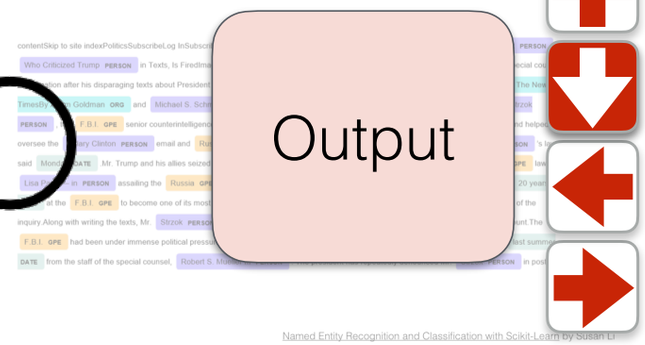
Wu *et al.* Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation

Word Embeddings



On word embeddings - Part 1 by Ruder

Named Entity Recognition



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

- 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	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
#1 Single - nlnet	-	-	83.5	90
#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.+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

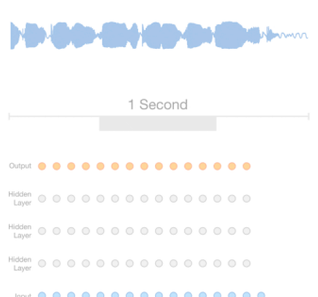
Speech Recognition

Word Error Rate

Senone set	Model/combination step	WER devset ngram-LM	WER test ngram-LM	WER devset LSTM-LMs	WER test LSTM-LMs
9k	BLSTM	11.5	8.3	9.2	6.3
27k	BLSTM	11.4	8.0	9.3	6.3
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]	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 [30]	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**
[illegible]	[illegible]	[illegible]	[illegible]	80.9	95.6
PolyNet [69]	331×331	92 M	34.7 B	81.3	95.8
DPN-131 [8]	320×320	80 M	32.0 B	81.5	95.8
SENet [25]	**320×320**	**115 M**	**42.3 B**	**82.7**	**96.2**
NASNet-A (6 @ 4032)	**331×331**	**88.9 M**	**23.8 B**	**82.7**	**96.2**

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

Zoph *et al.* Learning Transferable Architectures for Scalable Image Recognition

Instance Segmentation

Figure 5. Mask R-CNN 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

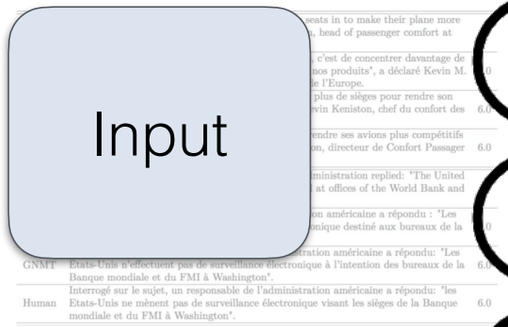
Visual Question Answering

Method	VQA v2 test-dev All	Yes/no	Num.	Other	VQA v2 test-std All	Yes/no	Num.	Other
Prior (most common answer in training set) [14]	–	–	–	–	25.98	61.20	0.36	1.17
LSTM Language only (blind model) [14]	–	–	–	–	44.26	67.01	31.55	27.37
Deeper LSTM Q norm. I [7] as reported in [14]	–	–	–	–	54.22	73.46	35.18	41.83
MCB [7] as reported in [14]	–	–	–	–	62.27	78.82	38.28	53.36
UPMC-LIP6 [7]	–	–	–	–	65.71	82.07	41.06	57.12
Athena	–	–	–	–	67.59	82.50	44.19	59.97
LV-NUS	–	–	–	–	66.77	81.89	46.29	58.30
HDU-USYD-UNCC	–	–	–	–	68.09	84.50	45.39	59.01
Proposed model								
ResNet features 7×7, single network	62.07	79.20	39.46	52.62	62.27	79.32	39.77	52.59
Image features from bottom-up attention, adaptive K, single network	65.32	81.82	44.21	56.05	65.67	82.20	43.90	56.26
ResNet features 7×7, ensemble	66.34	83.38	43.17	57.10	66.73	83.71	43.77	57.20
Image features from bottom-up attention, adaptive K, ensemble	**69.87**	**86.08**	**48.99**	**60.80**	**70.34**	**86.60**	**48.64**	**61.15**

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

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

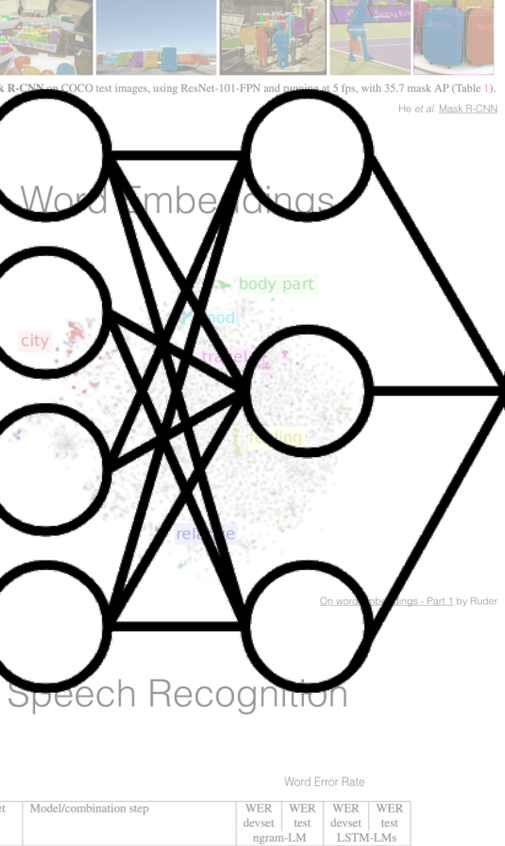
Machine Translation



	[illegible]eats in to make their plane more [illegible], head of passenger comfort at
	[illegible] c'est de concentrer davantage de [illegible]nos produits', a déclaré Kevin M. [illegible]e l'Europe.
	[illegible] plus de sièges pour rendre son [illegible]evin Keniston, chef du confort des
	[illegible]rendre ses avions plus compétitifs [illegible]on, directeur de Confort Passager
	[illegible]ministration replied: "The United [illegible] at offices of the World Bank and
	[illegible]on américaine a répondu : "Les [illegible]nique destiné aux bureaux de la
GNMT	[illegible]tration américaine a répondu: "Les Etats-Unis n'effectuent pas de surveillance électronique à l'intention des bureaux de la Banque mondiale et du FMI à Washington".
Human	Interrogé sur le sujet, un responsable de l'administration américaine a répondu: "les Etats-Unis ne mènent pas de surveillance électronique visant les sièges de la Banque mondiale et du FMI à Washington".

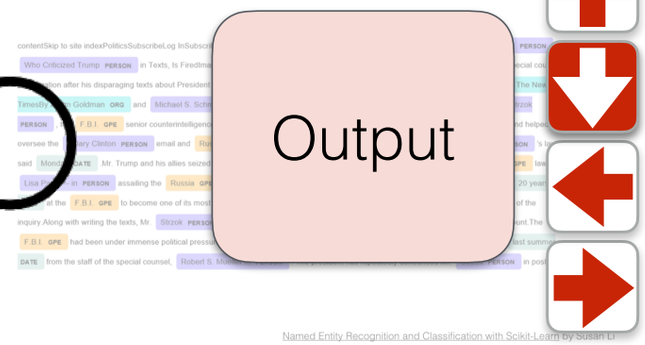
Wu *et al.* Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation

Word Embeddings



On word embeddings - Part 1 by Ruder

Named Entity Recognition



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

- 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	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
#1 Single - nlnet	-	-	83.5	90.1
#2 Single - QANet	-	-	82.5	89.3
Published				
BiDAF+ELMo (Single)	-	85.8	-	-
R.M. Reader (Single)	78.9	86.3	79.5	86.6
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours				
$BERT_{BASE}$ (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**
$BERT_{LARGE}$ (Ens.+TriviaQA)	**86.2**	**92.2**	**87.4**	**93.2**

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

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

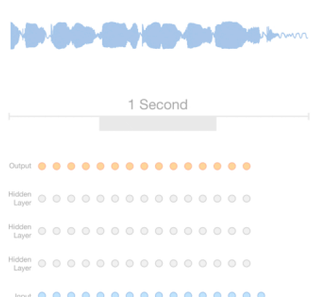
Speech Recognition

Word Error Rate

Senone set	Model/combination step	WER devset ngram-LM	WER test ngram-LM	WER devset LSTM-LMs	WER test LSTM-LMs
9k	BLSTM	11.5	8.3	9.2	6.3
27k	BLSTM	11.4	8.0	9.3	6.3
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]	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 [30]	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**
[illegible]	[illegible]	[illegible]	[illegible]	80.9	95.6
PolyNet [69]	331×331	92 M	34.7 B	81.3	95.8
DPN-131 [8]	320×320	80 M	32.0 B	81.5	95.8
SENet [25]	**320×320**	**115 M**	**42.3 B**	**82.7**	**96.2**
NASNet-A (6 @ 4032)	**331×331**	**88.9 M**	**23.8 B**	**82.7**	**96.2**

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

Zoph *et al.* Learning Transferable Architectures for Scalable Image Recognition

Instance Segmentation

Fig. 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

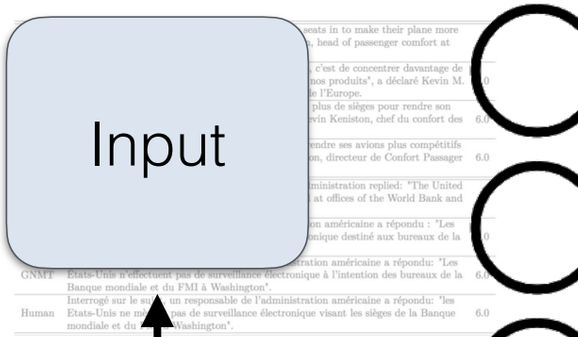
Visual Question Answering

Method	VQA v2 test-dev All	Yes/no	Num.	Other	VQA v2 test-std All	Yes/no	Num.	Other
Prior (most common answer in training set) [14]	–	–	–	–	25.98	61.20	0.36	1.17
LSTM Language only (blind model) [14]	–	–	–	–	44.26	67.01	31.55	27.37
Deeper LSTM Q norm. I [7] as reported in [14]	–	–	–	–	54.22	73.46	35.18	41.83
MCB [7] as reported in [14]	–	–	–	–	62.27	78.82	38.28	53.36
UPMC-LIP6 [7]	–	–	–	–	65.71	82.07	41.06	57.12
Athena	–	–	–	–	67.59	82.50	44.19	59.97
LV-NUS	–	–	–	–	66.77	81.89	46.29	58.30
HDU-USYD-UNCC	–	–	–	–	68.09	84.50	45.39	59.01
Proposed model								
ResNet features 7×7, single network	62.07	79.20	39.46	52.62	62.27	79.32	39.77	52.59
Image features from bottom-up attention, adaptive K, single network	65.32	81.82	44.21	56.05	65.67	82.20	43.90	56.26
ResNet features 7×7, ensemble	66.34	83.38	43.17	57.10	66.73	83.71	43.77	57.20
Image features from bottom-up attention, adaptive K, ensemble	**69.87**	**86.08**	**48.99**	**60.80**	**70.34**	**86.60**	**48.64**	**61.15**

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

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

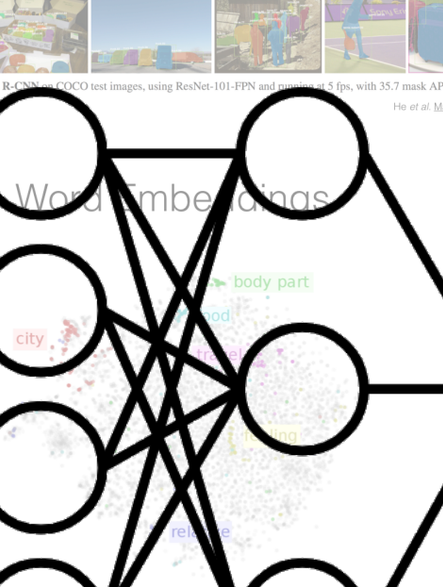
Machine Translation



PBMT	[illegible]eats in to make their plane more [illegible], head of passenger comfort at
GNMT	[illegible] c'est de concentrer davantage de [illegible]nos produits', a déclaré Kevin M. [illegible] de l'Europe.
Human	[illegible] plus de sièges pour rendre son [illegible]evin Keniston, chef du confort des
PBMT	[illegible]endre ses avions plus compétitifs [illegible]on, directeur de Confort Passager
GNMT	[illegible]ministration replied: "The United [illegible] at offices of the World Bank and
Human	[illegible]on américaine a répondu : "Les [illegible]nique destiné aux bureaux de la
PBMT	[illegible]ration américaine a répondu: "Les [illegible] Etats-Unis n'effectuent pas de surveillance électronique à l'intention des bureaux de la Banque mondiale et du FMI à Washington".
GNMT	Interrogé sur le su[illegible], un responsable de l'administration américaine a répondu: "les Etats-Unis ne m[illegible]pas de surveillance électronique visant les sièges de la Banque mondiale et du [illegible]Washington".
Human	

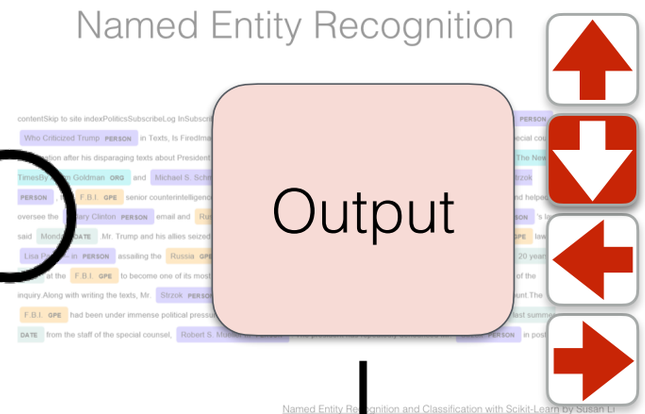
Wu *et al.* Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation

Word Embeddings



On word embeddings - Part 1 by Ruder

Named Entity Recognition



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

- 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	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
#1 Single - nlnet	-	-	83.5	90.1
#2 Single - QANet	-	-	82.5	89.3
Published				
BiDAF+ELMo (Single)	-	85.8	-	-
R.M. Reader (Single)	78.9	86.3	79.5	86.6
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours				
$BERT_{BASE}$ (Single)	80.8	88.5	-	-
[illegible] (Single)	84.1	90.9	-	-
$BERT_{LARGE}$ (Ensemble)	85.8	91.8	-	-
$BERT_{LARGE}$ (Sgl.+TriviaQA)	**84.2**	**91.1**	**85.1**	**91.8**
$BERT_{LARGE}$ (Ens.+TriviaQA)	**86.2**	**92.2**	**87.4**	**93.2**

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

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

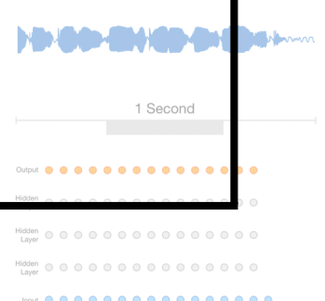
Speech Recognition

Word Error Rate

Senone set	Model/combination step	WER devset ngram-LM	WER test ngram-LM	WER devset LSTM-LMs	WER test LSTM-LMs
9k	BLSTM	11.5	8.3	9.2	6.3
27k	BLSTM	11.4	8.0	9.3	6.3
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

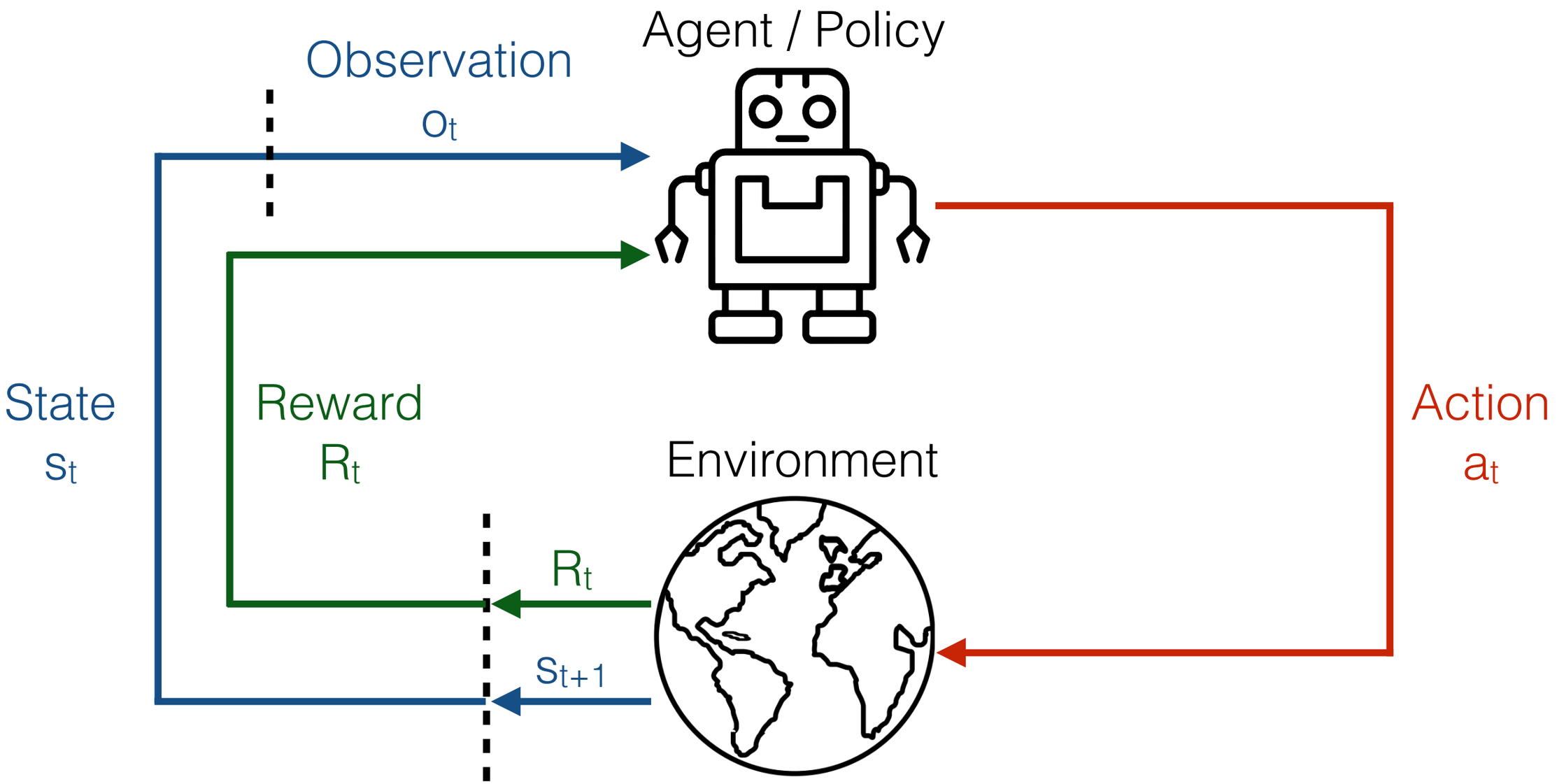
Xiong *et al.* The Microsoft 2017 Conversational Speech Recognition System

Speech Synthesis (text-to-speech)



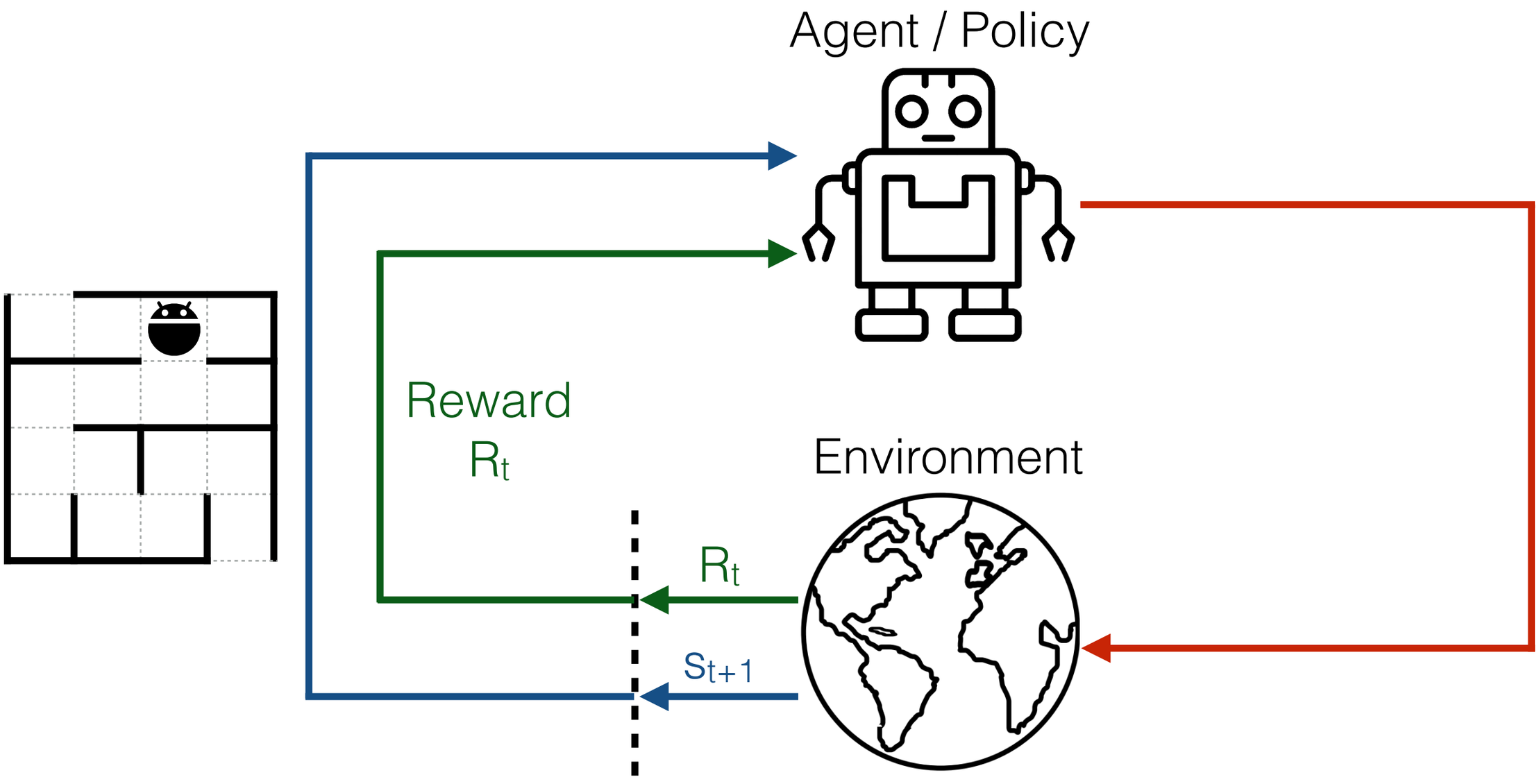
Van Den Oord *et al.* WaveNet: A Generative Model for Raw Audio

Robot Learning via Reinforcement Learning



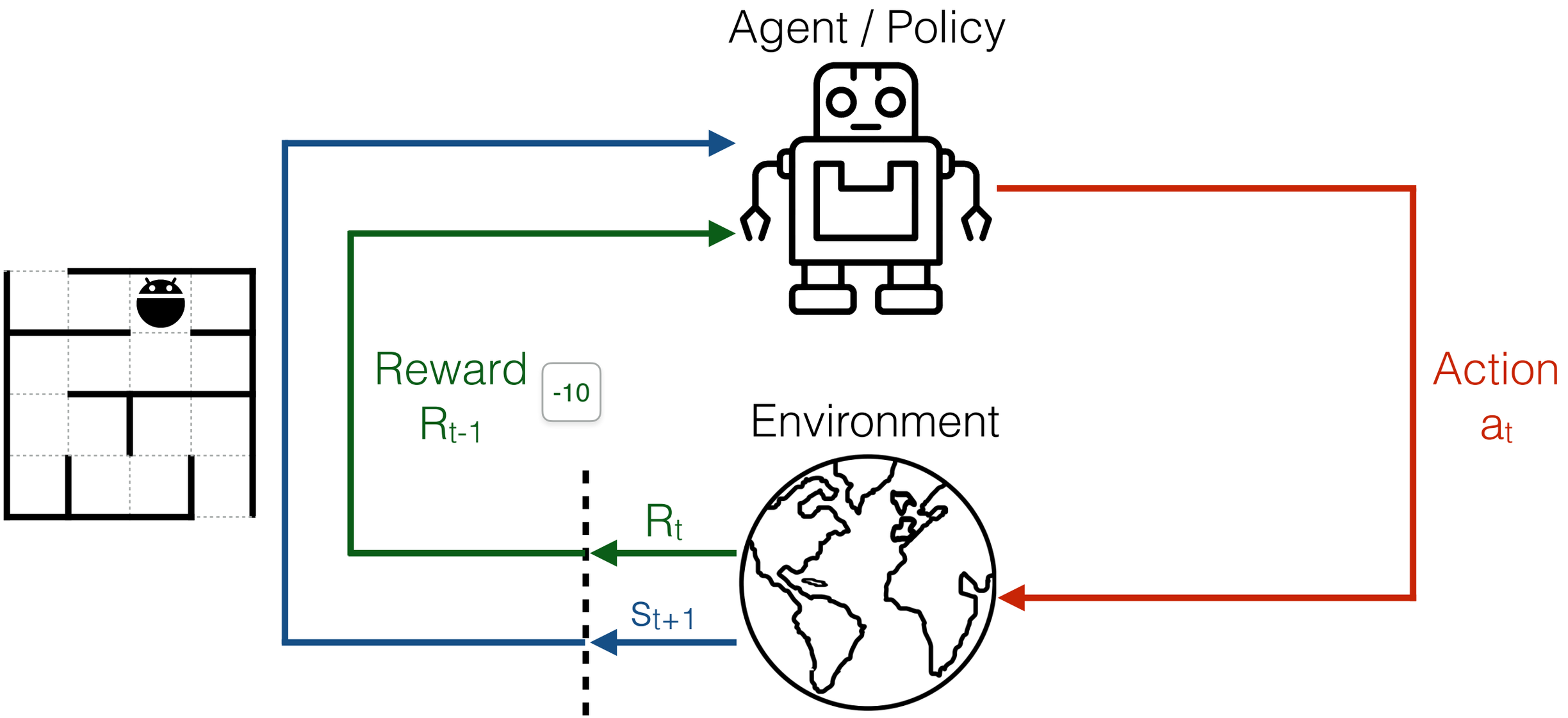
Goal: maximize $\sum_{t=0}^{t=H} \gamma^t R_t(s_t, a_t)$

Robot Learning via Reinforcement Learning



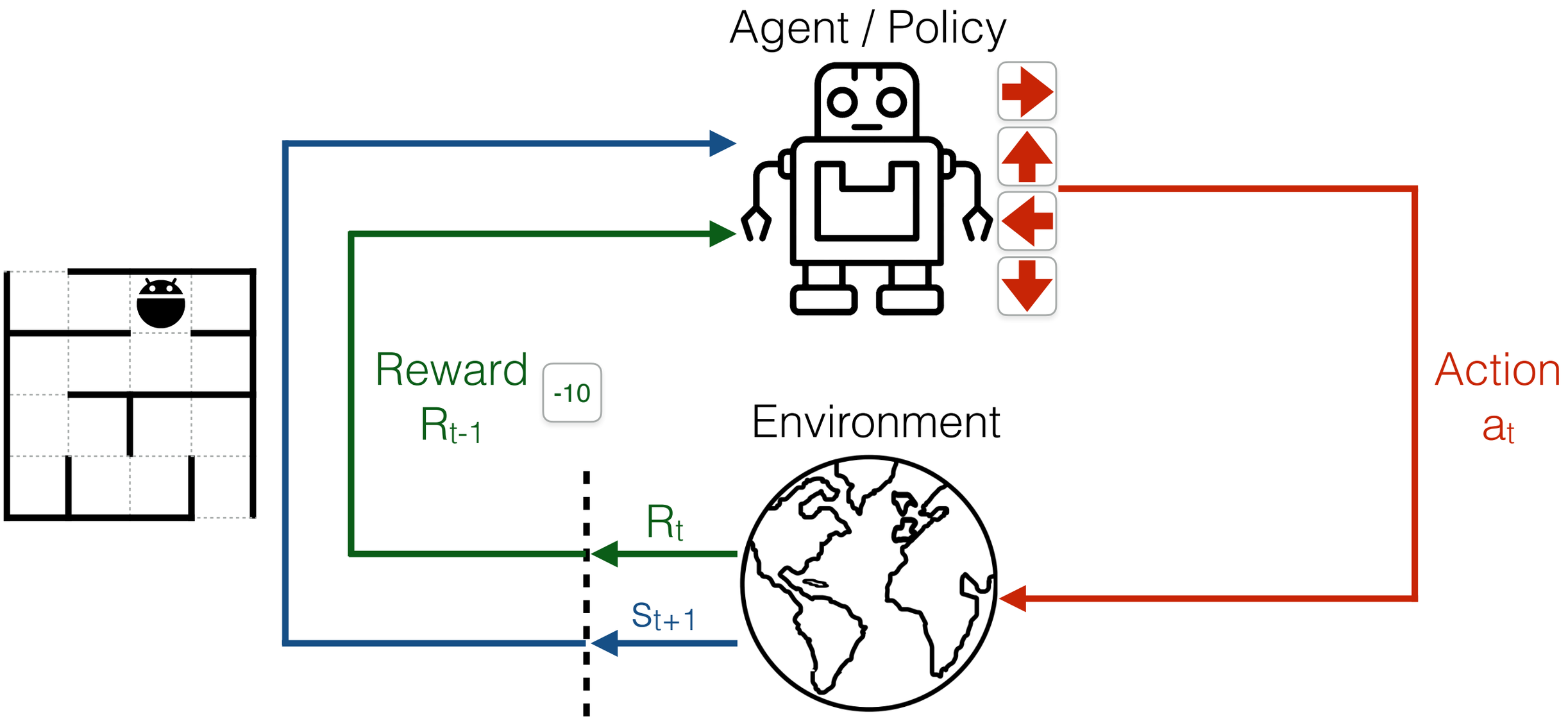
Goal: maximize $\sum_{t=0}^{t=H} \gamma^t R_t(s_t, a_t)$

Robot Learning via Reinforcement Learning



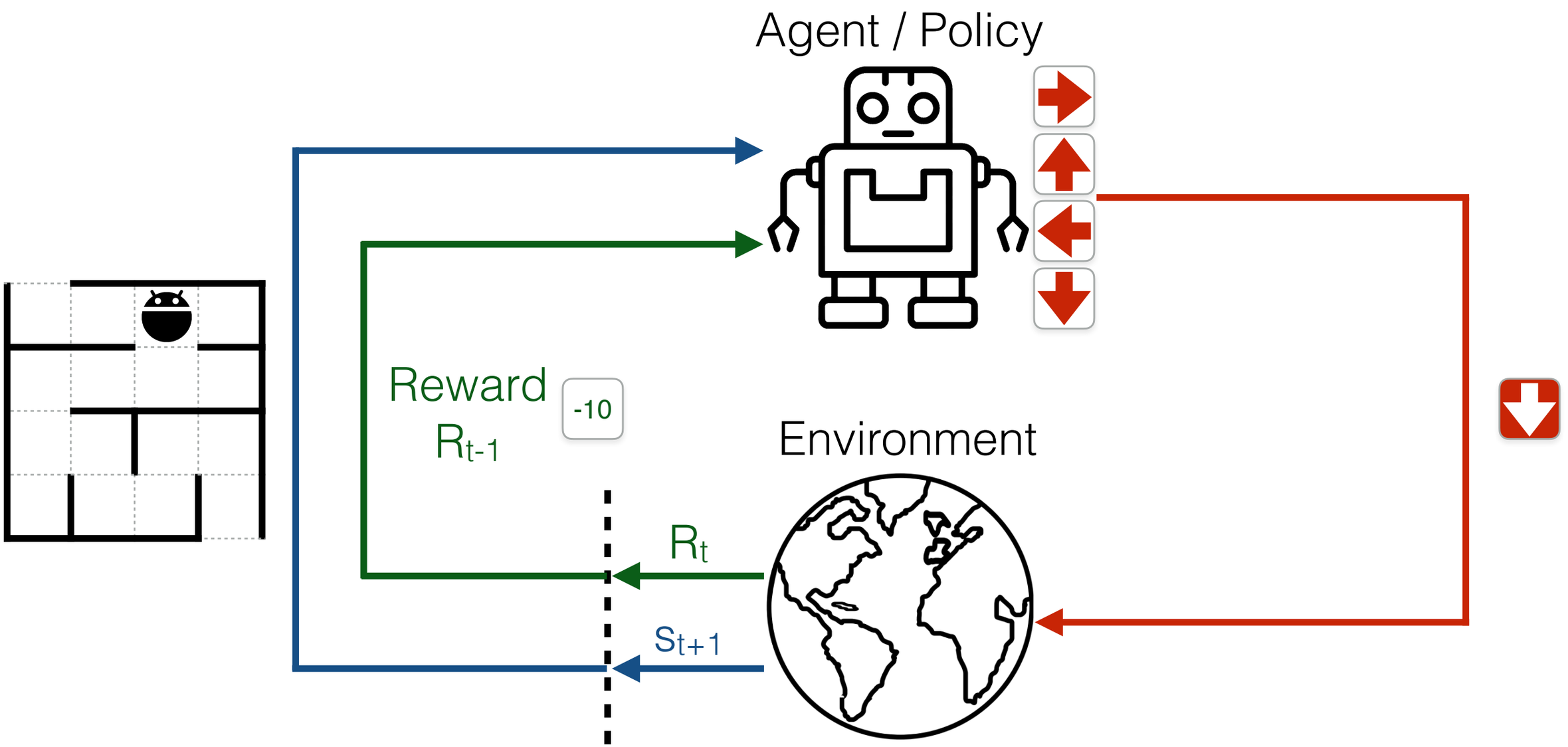
Goal: maximize $\sum_{t=0}^{t=H} \gamma^t R_t(s_t, a_t)$

Robot Learning via Reinforcement Learning



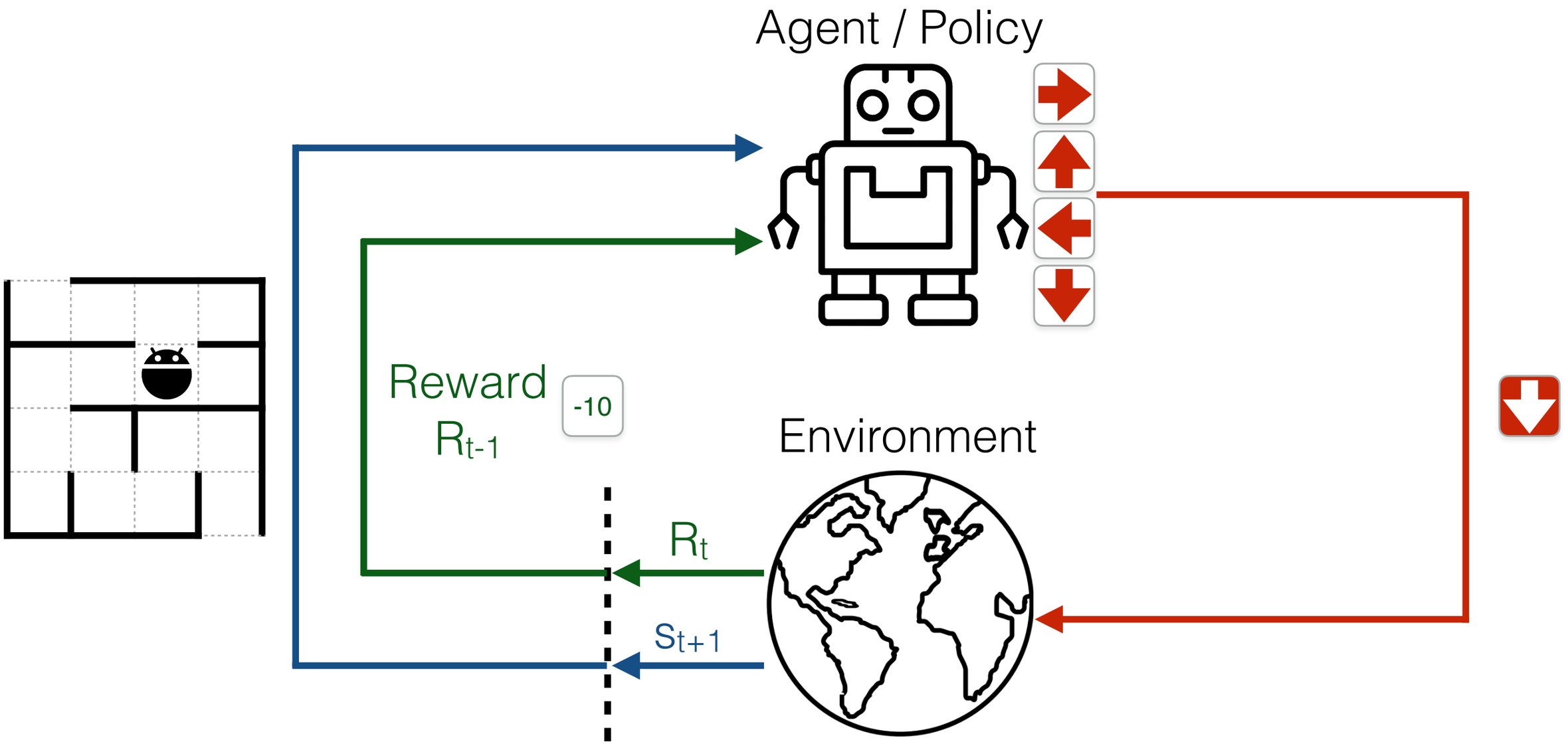
Goal: maximize $\sum_{t=0}^{t=H} \gamma^t R_t(s_t, a_t)$

Robot Learning via Reinforcement Learning



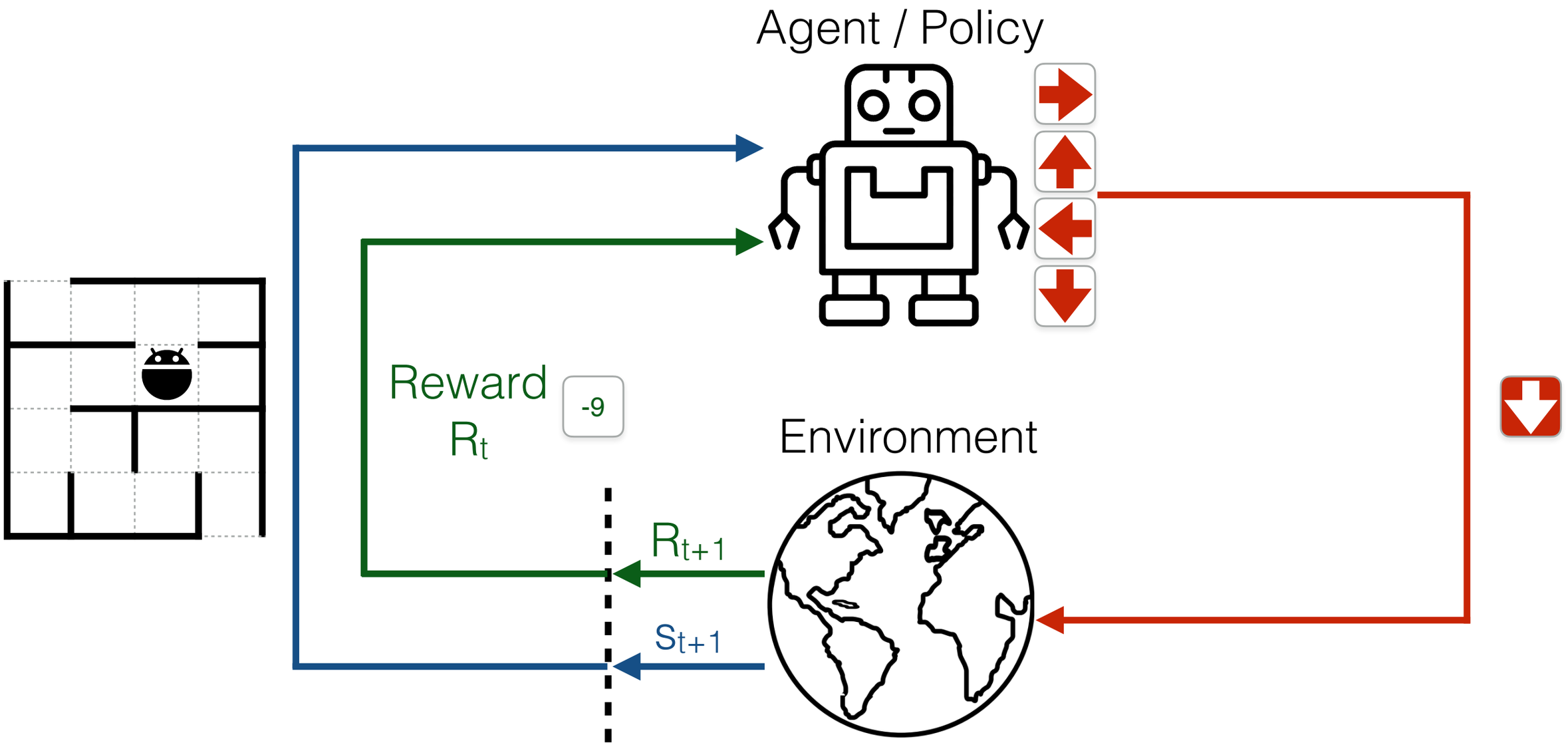
Goal: maximize $\sum_{t=0}^{t=H} \gamma^t R_t(s_t, a_t)$

Robot Learning via Reinforcement Learning



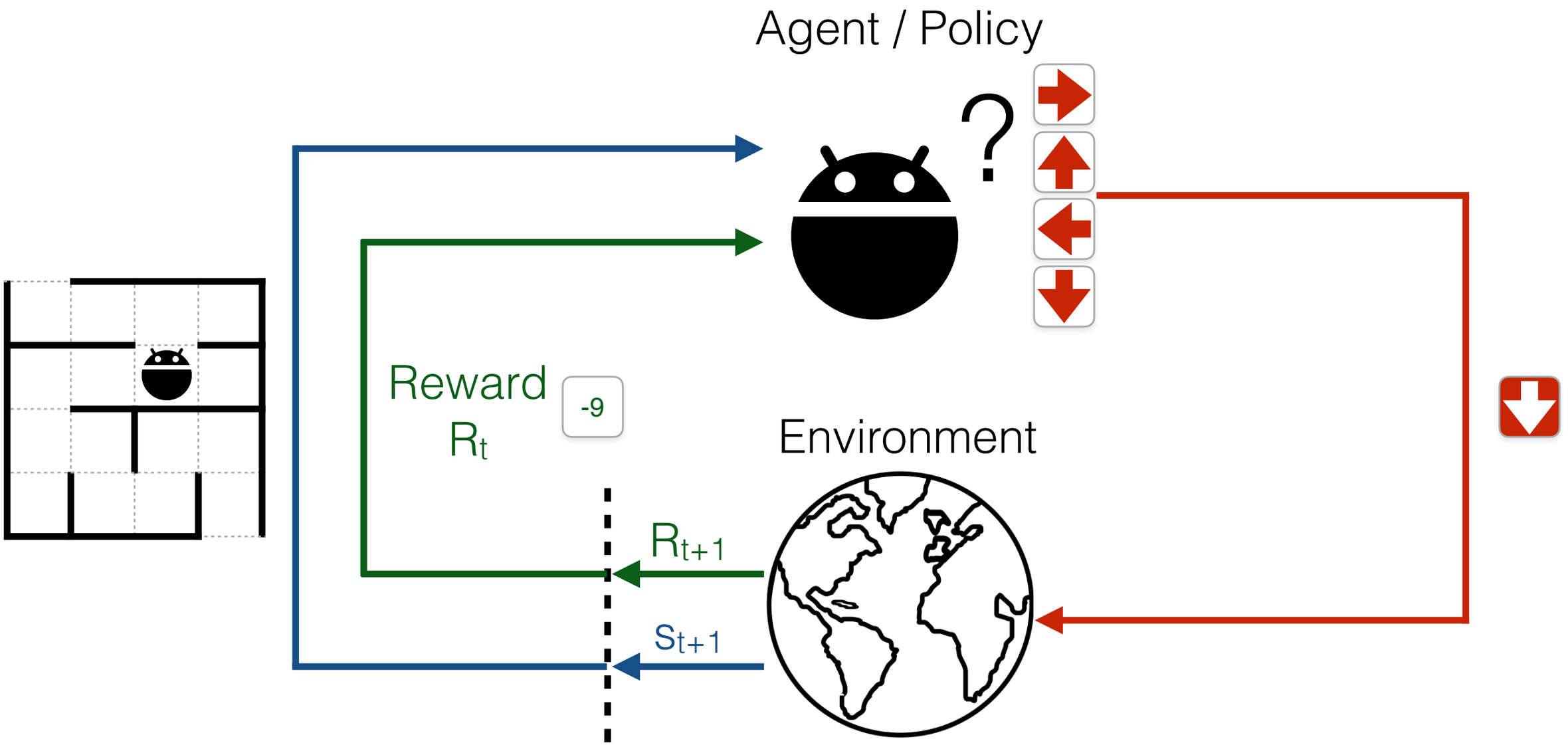
Goal: maximize $\sum_{t=0}^{t=H} \gamma^t R_t(s_t, a_t)$

Robot Learning via Reinforcement Learning



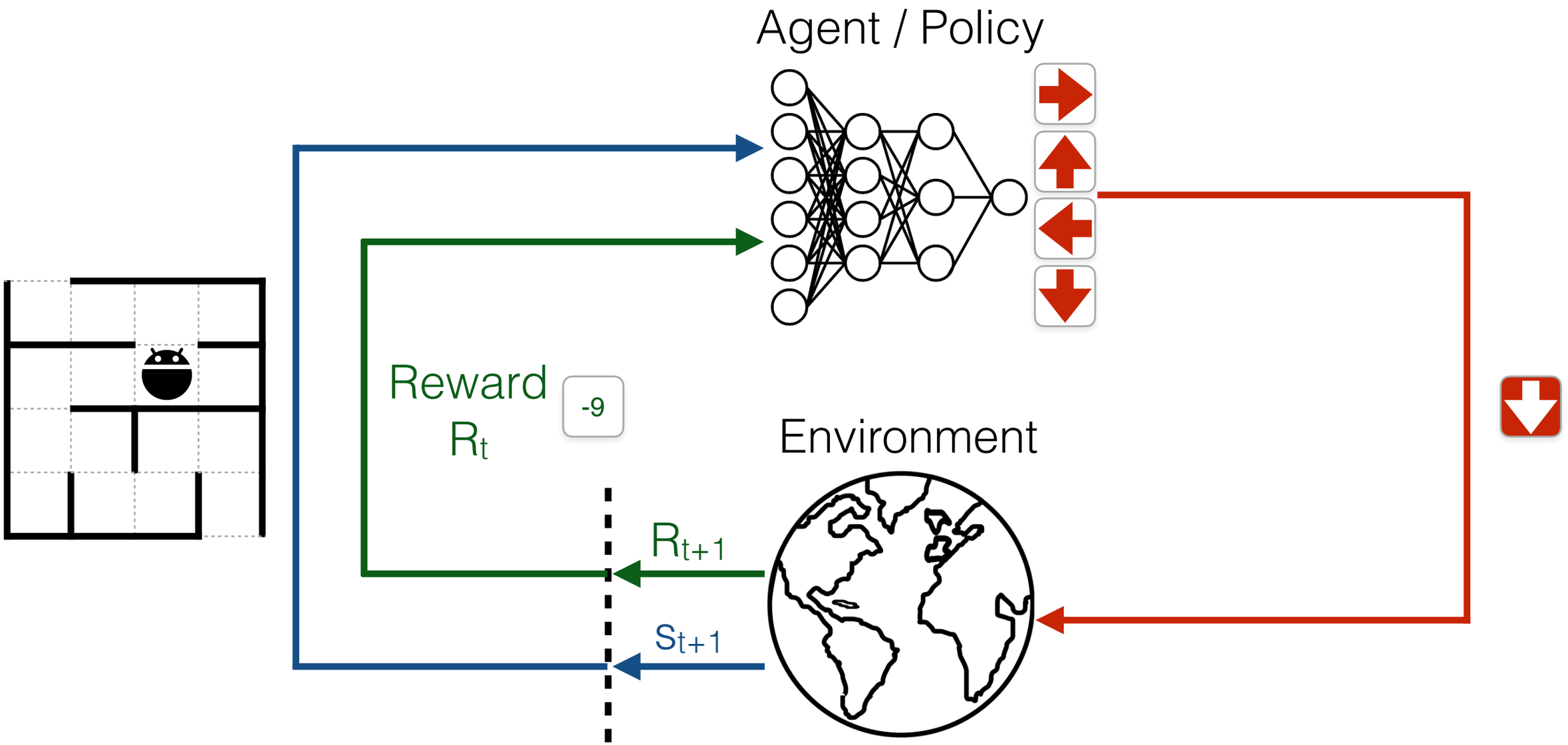
Goal: maximize $\sum_{t=0}^{t=H} \gamma^t R_t(s_t, a_t)$

Robot Learning via Reinforcement Learning



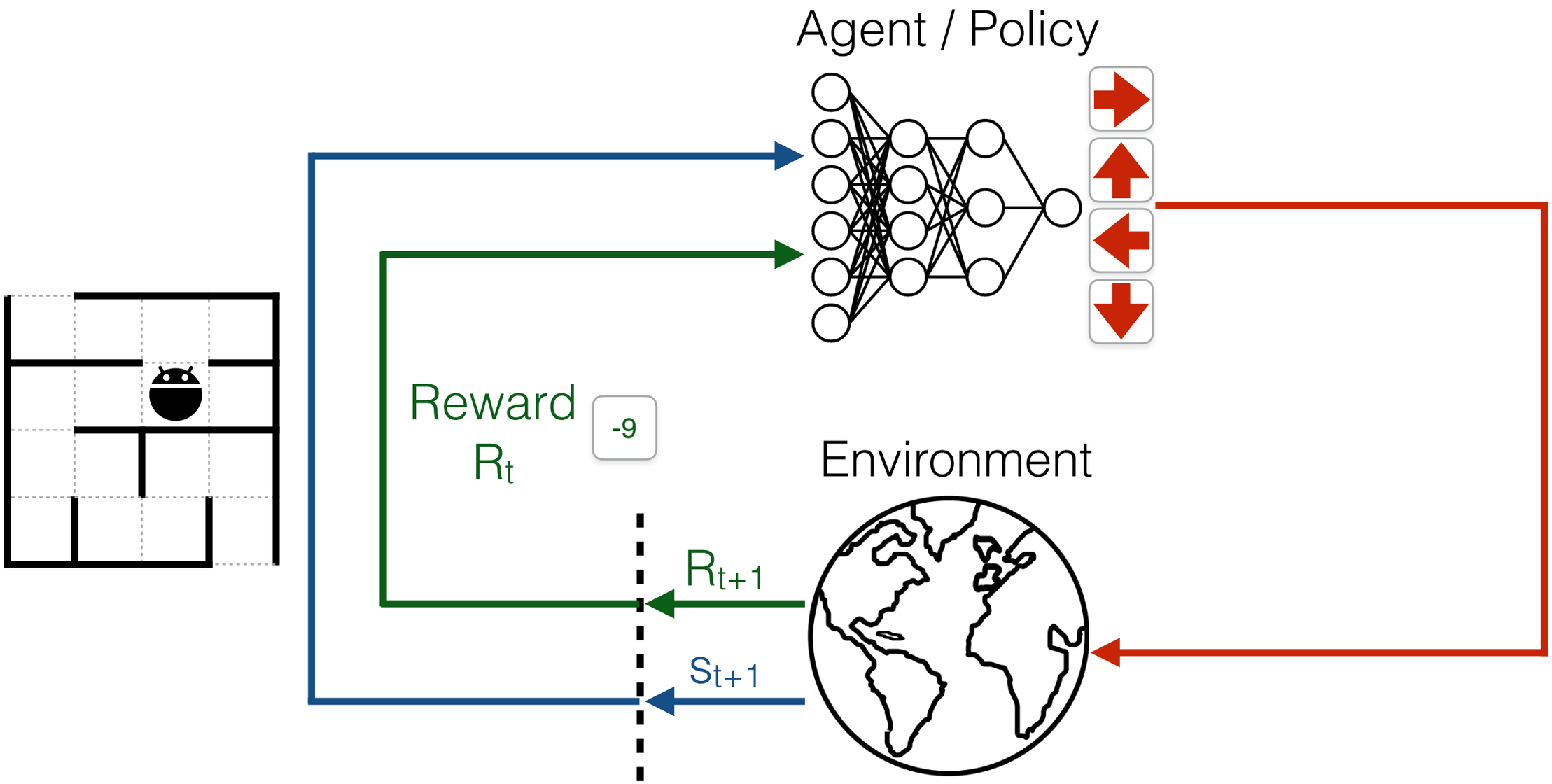
Goal: maximize $\sum_{t=0}^{t=H} \gamma^t R_t(s_t, a_t)$

Robot Learning via Deep Reinforcement Learning



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

Robot Learning via Deep Reinforcement Learning



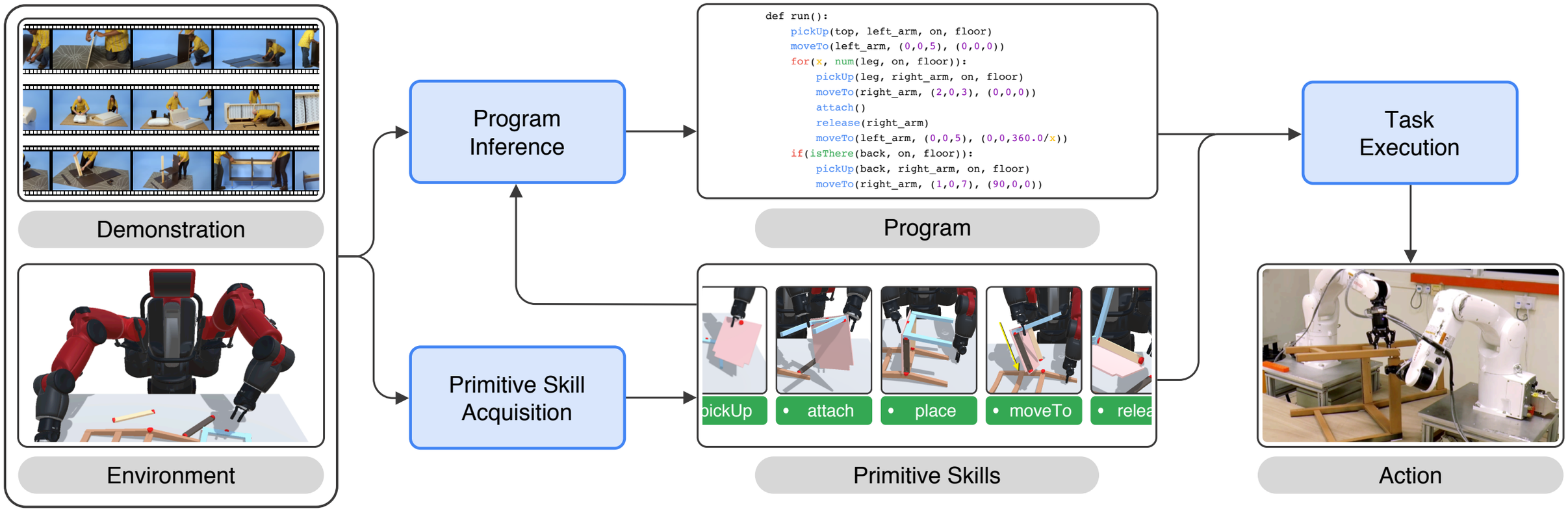
Uninterpretable

Not generalizable

Limited to short-horizon tasks

No skill-reuse

Program-Guided Framework for Interpreting and Acquiring Complex Skills



Interpretable

Programmatic / Generalizable

Hierarchical

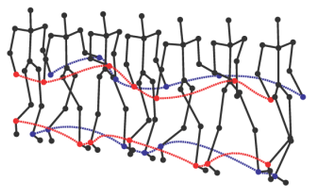
Modular

Program-Guided Framework for Interpreting and Acquiring Complex Skills

Skill Specification → Program → High-level Plan → Low-level Execution

Video demonstration

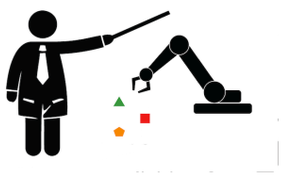
Trajectory demonstration



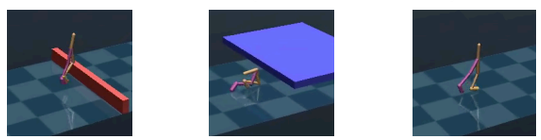
Language instruction

Stir-fry the onions until tender, and repeat this for garlic and carrots, if you have soy sauce, add some. Pour 2/3 cups the whisked eggs into the stir-fried and scramble.

Reward



Actions (sub-skills)



`Jump` `Crawl` `Walk`

Perceptions

`isCurb()` `isCeiling()`

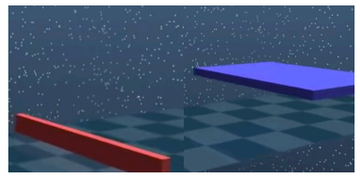
Control Flows

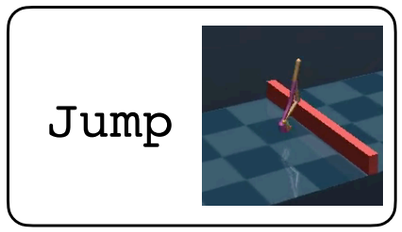
`while` `if` `else` `elseIf`

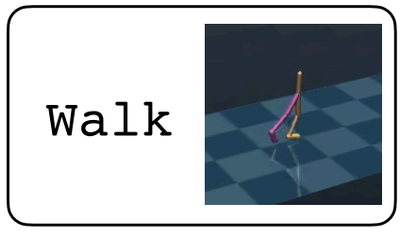
Program

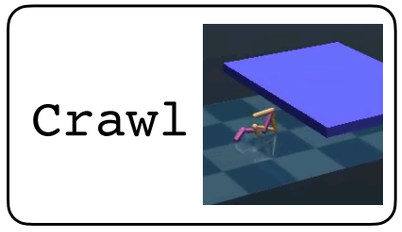
```
DEF run()
    while True:
        if isCurb():
            jump
        elseIf isCeiling():
            crawl()
        else:
            walk()
```

Observation

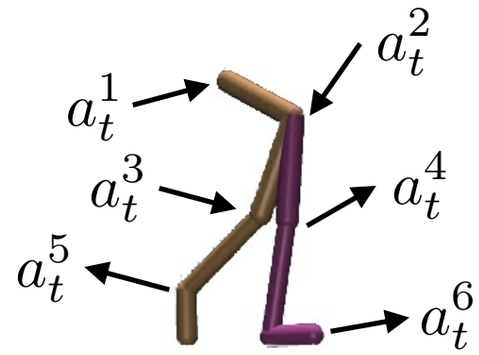








Low-level execution

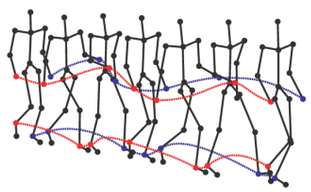


Program Inference

Skill Specification → Program → High-level Plan → Low-level Execution

Video demonstration

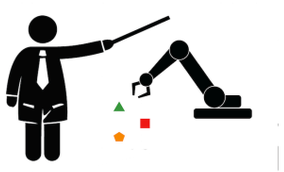
Trajectory demonstration



Language instruction

Stir-fry the onions until tender, and repeat this for garlic and carrots, if you have soy sauce, add some. Pour 2/3 cups the whisked eggs into the stir-fried and scramble.

Reward



Actions (sub-skills)

`Jump` `Crawl` `Walk`

Perceptions

`isCurb()` `isCeiling()`

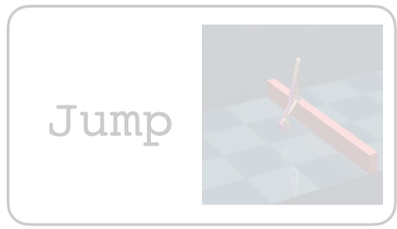
Control Flows

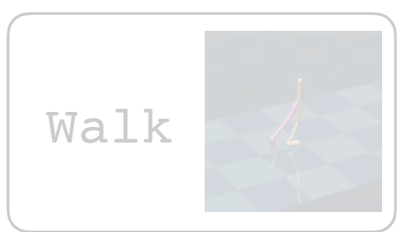
`while` `if` `else` `elseIf`

Program

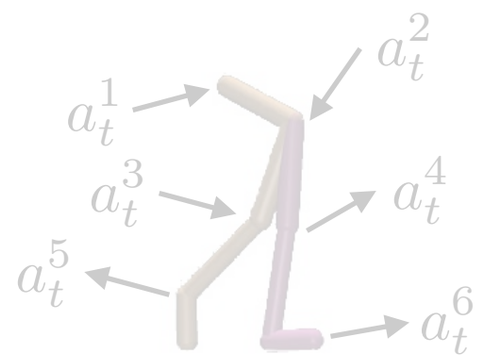
```
DEF run()
    while True:
        if isCurb():
            jump
        elseIf isCeiling():
            crawl()
        else:
            walk()
```

Observation



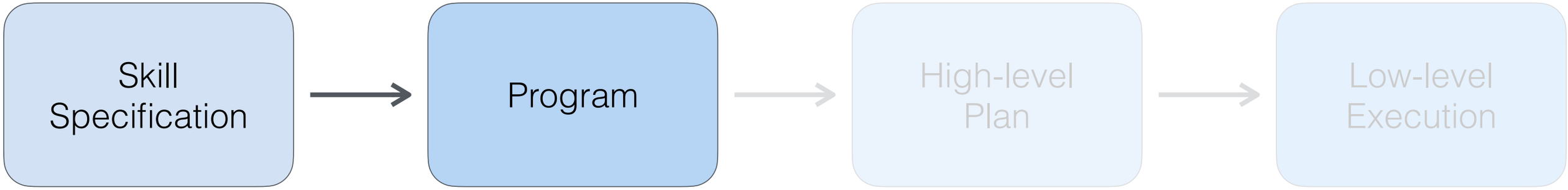


Low-level execution



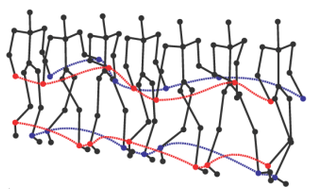
Neural Program Synthesis from Diverse Demonstration Videos

ICML 2018



Video demonstration

Trajectory demonstration



Program

```
DEF run()
    while True:
        if isCurb():
            jump
        elseIf isCeiling():
            crawl()
        else:
            walk()
```



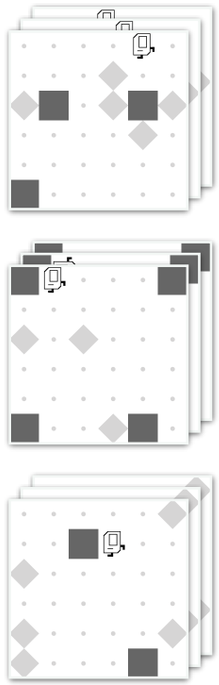
Hyeonwoo Noh


Sriram Somasundaram

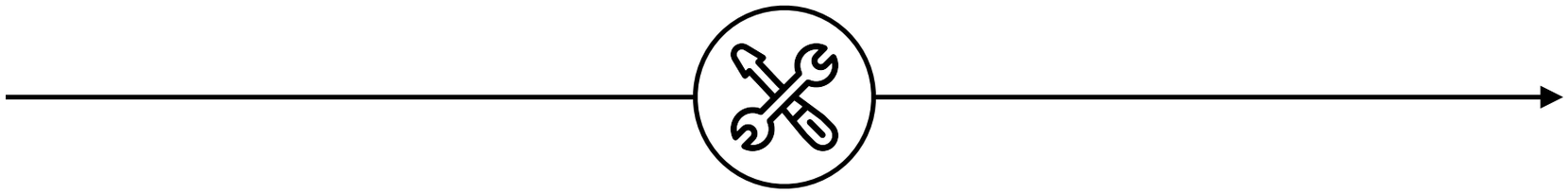

Joseph J. Lim

Imitation Learning

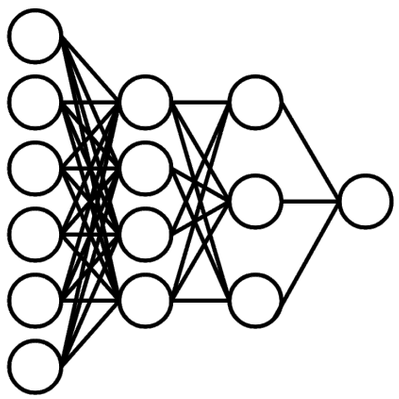
Demonstrations



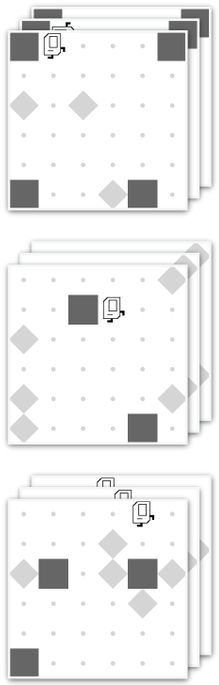
Imitate



Neural Network
Policy

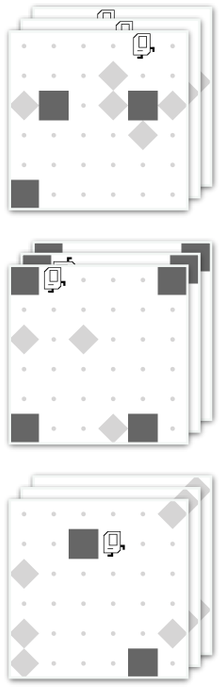


Execution



Imitation Learning by Synthesizing Programs

Demonstrations

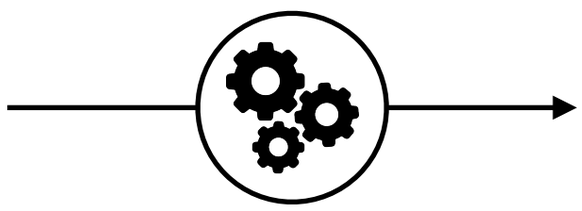


Synthesize

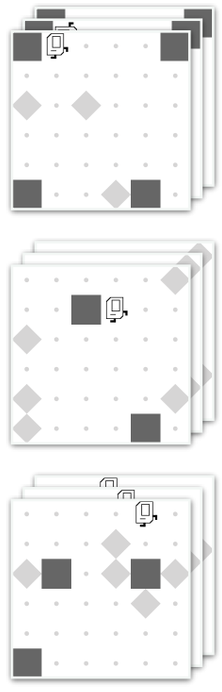
Program Policy

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

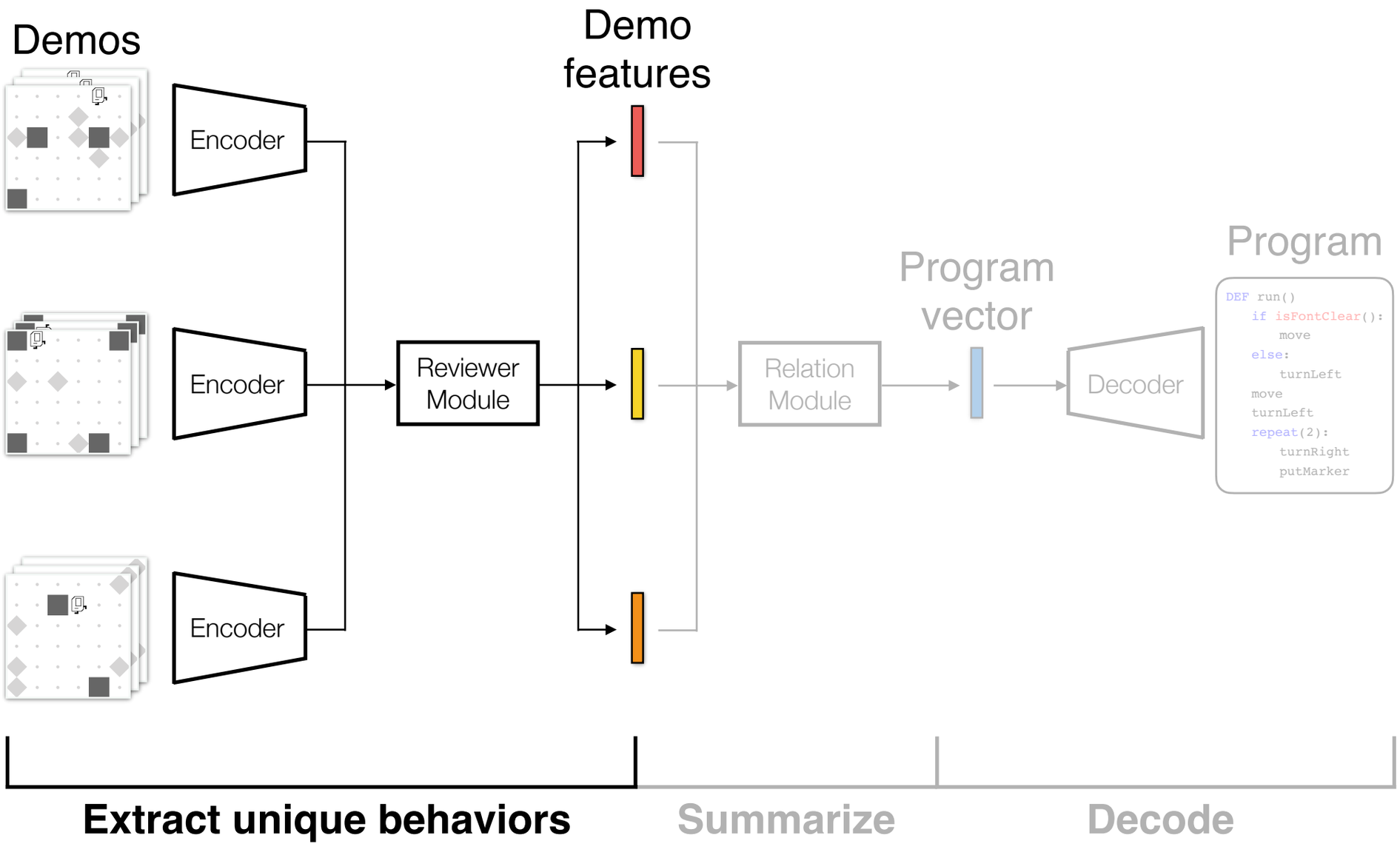
Execute



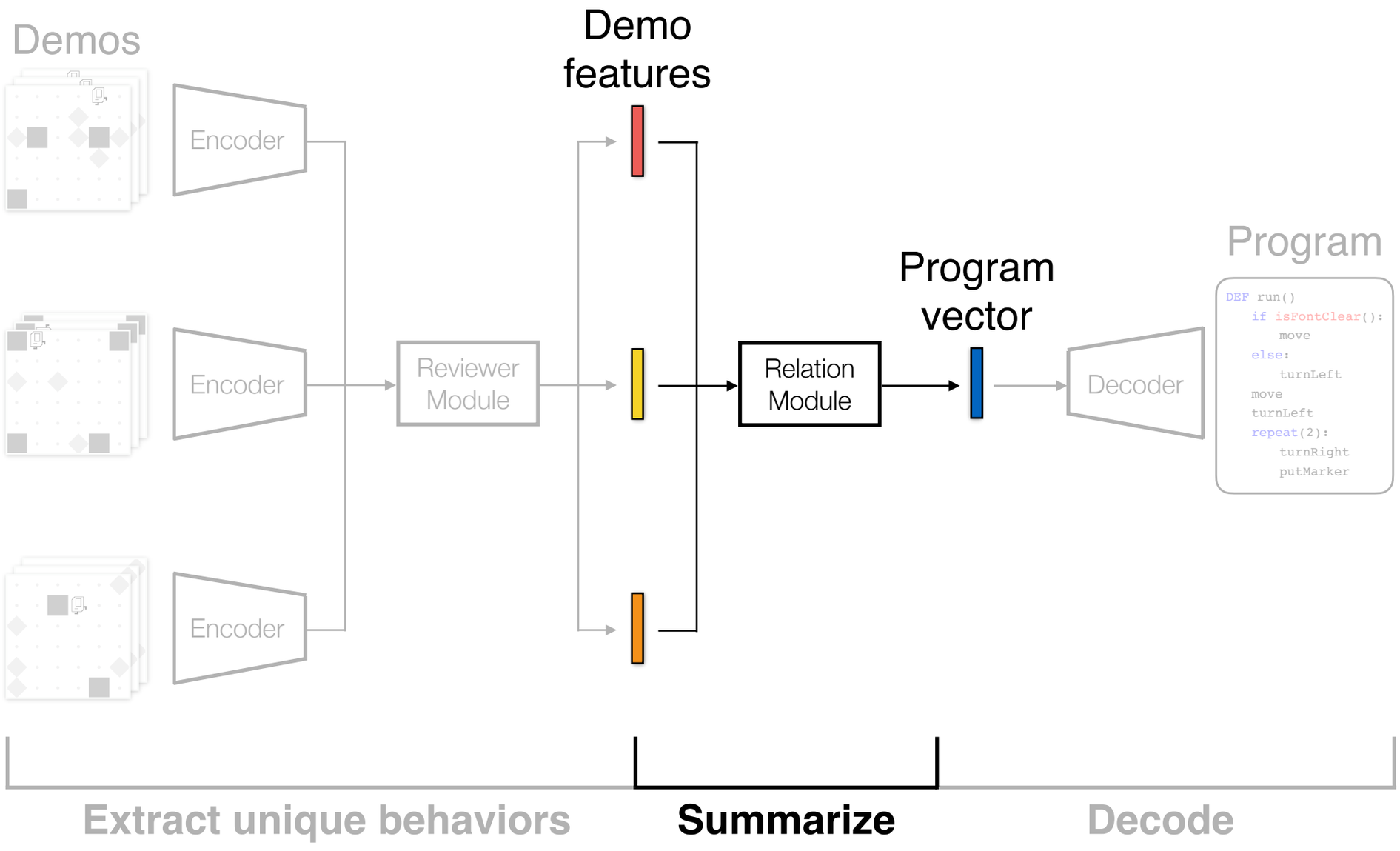
Execution



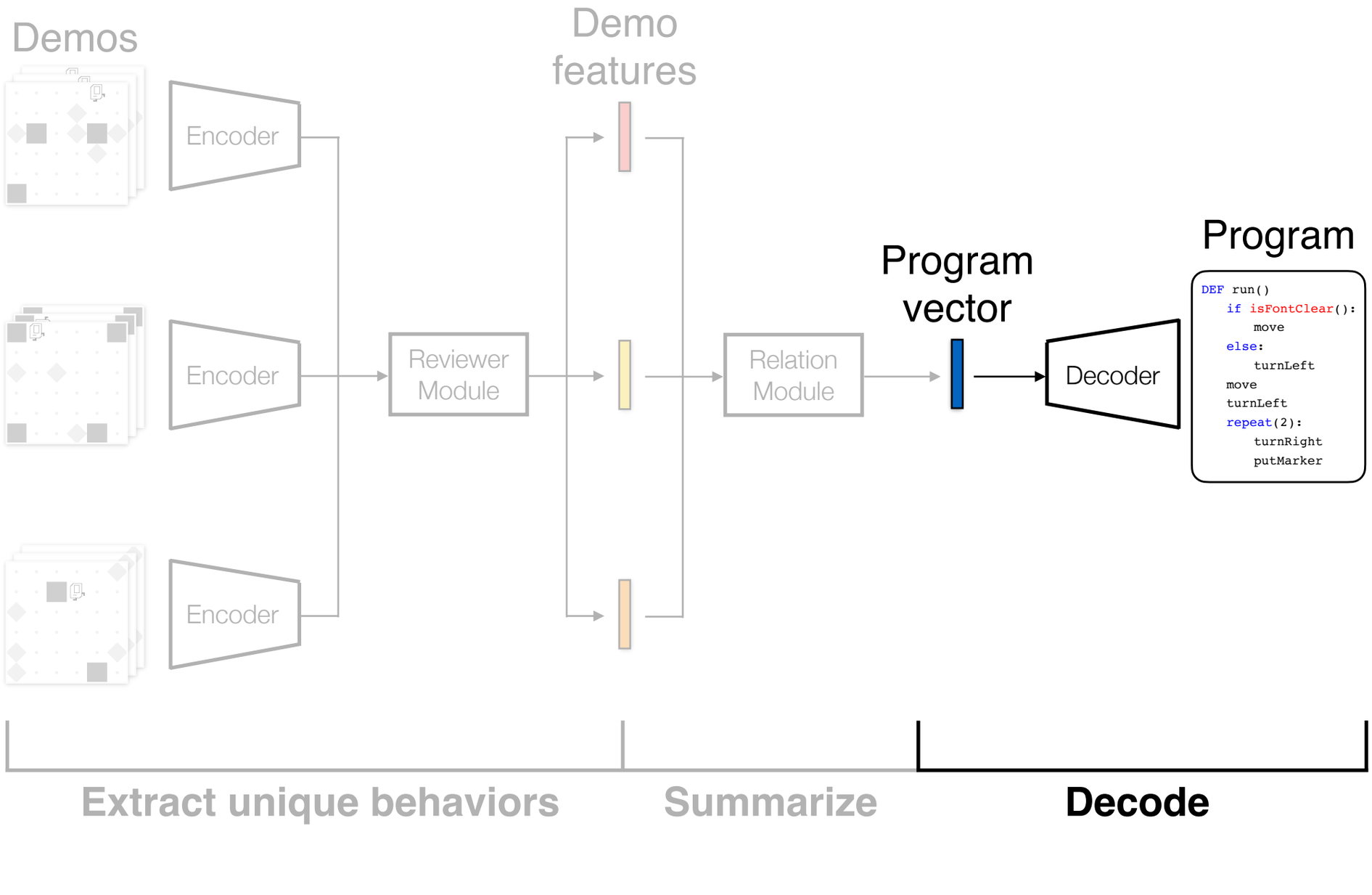
Model Overview



Model Overview



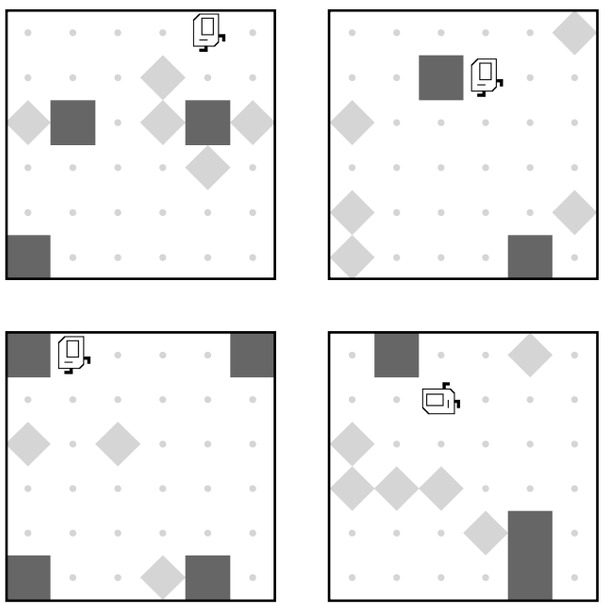
Model Overview



Environments

Karel

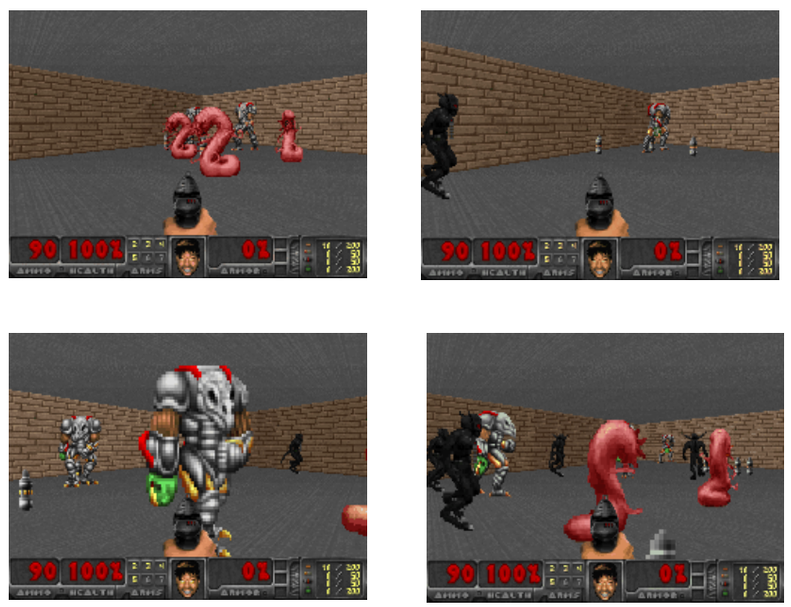
```
DEF run()
   if isFontClear():
      move
   else:
      turnLeft
   move
   turnLeft
   repeat(2):
      turnRight
      putMarker
```



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

ViZDoom

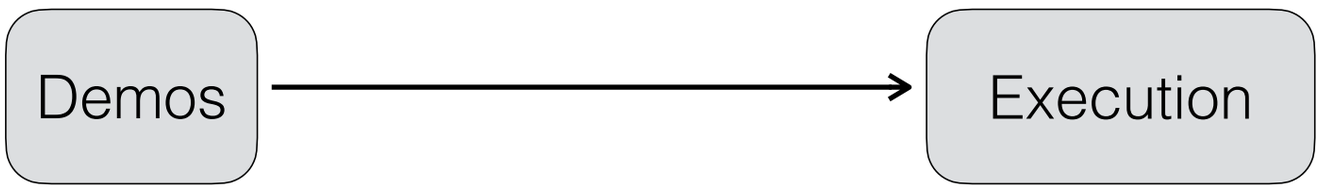
```
DEF run()
   while isFontClear(HellKnight):
      attack
      moveForward
   if isThere(Demon):
      moveRight
   else:
      moveLeft
      moveBackward
```



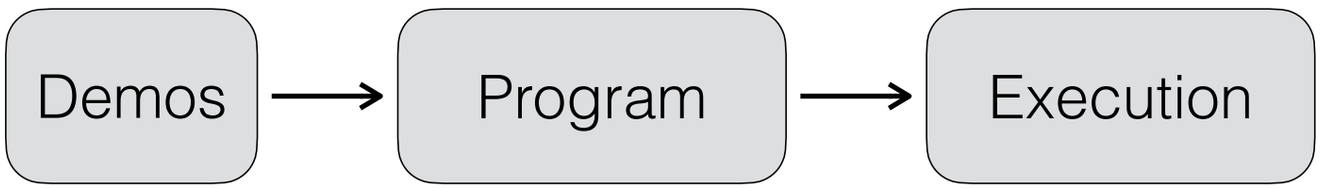
Kempka et al., “Vizdoom: A doom-based ai research platform for visual reinforcement learning.” In CIG, 2016

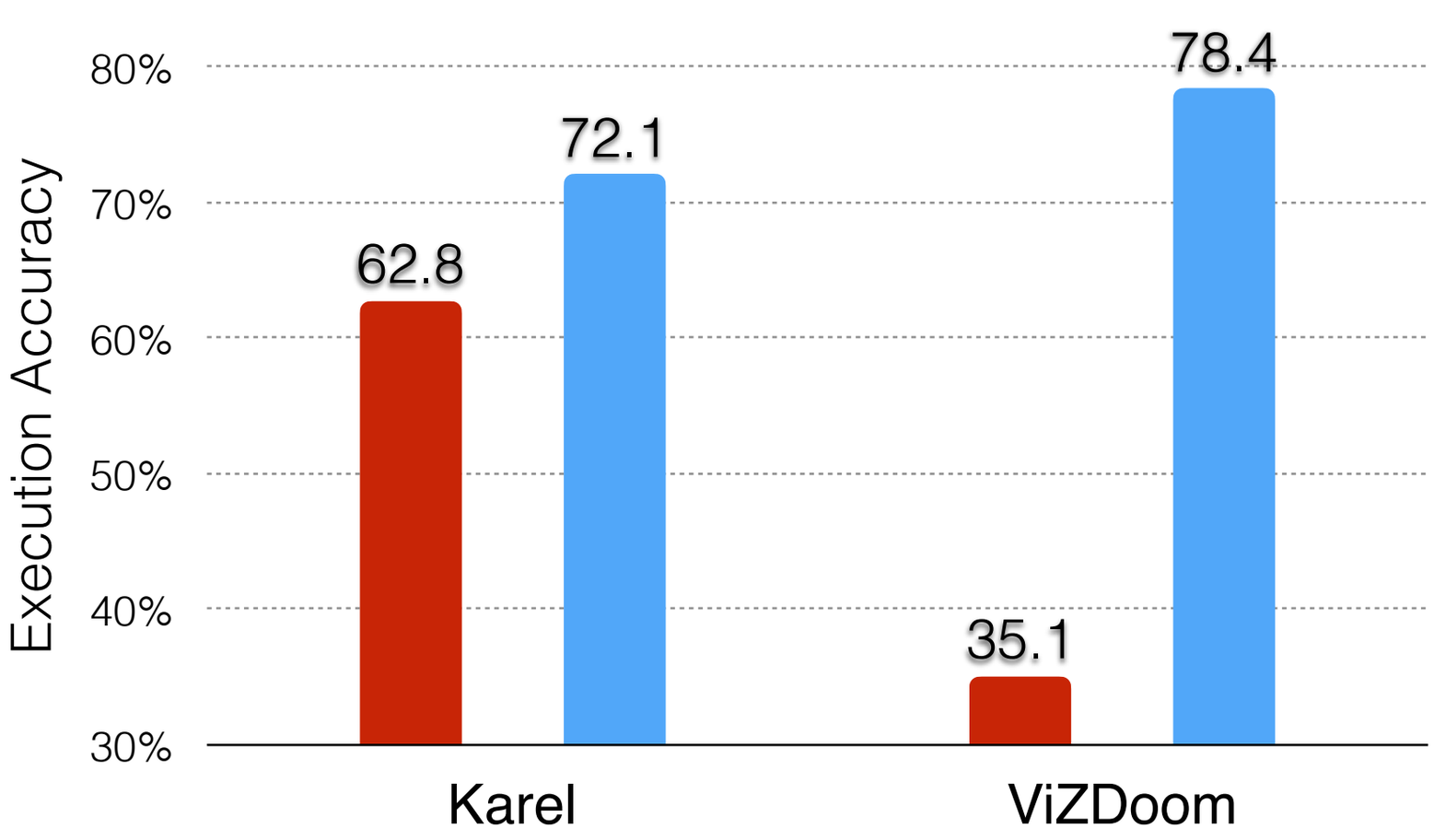
Quantitative Results

Neural Network Policy



Program Policy



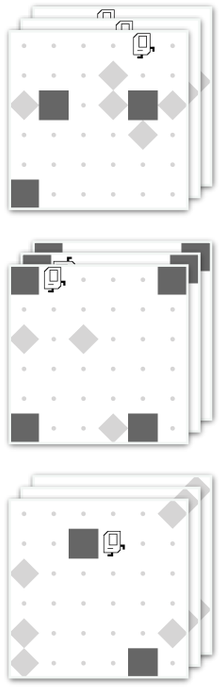



```
DEF run()
    if isFontClear():
        move
    else:
        turnLeft
    move
    turnLeft
    repeat(2):
        turnRight
        putMarker
```

Takeaway

- Synthesize **programs** to **imitate demonstrations**

Demonstrations

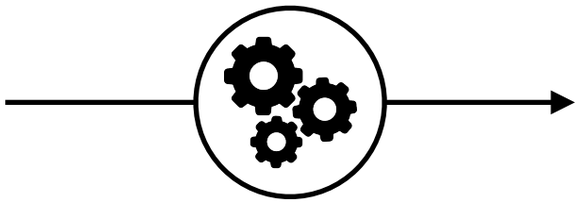


Synthesize

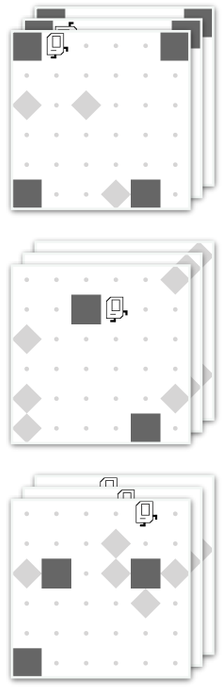
Program Policy

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

Execute

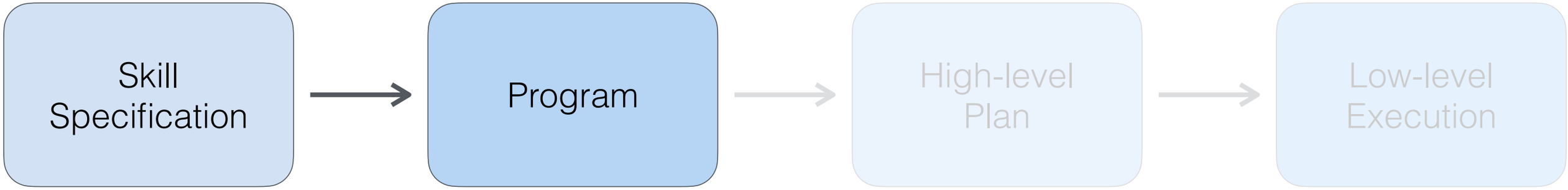


Execution

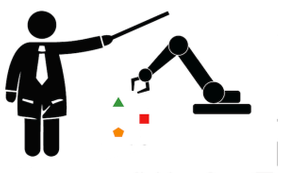


Learning to Synthesize Programs as Interpretable and Generalizable Policies

NeurIPS 2021



Reward



Program

```
DEF run()
    while True:
        if isCurb():
            jump
        elseIf isCeiling():
            crawl()
        else:
            walk()
```

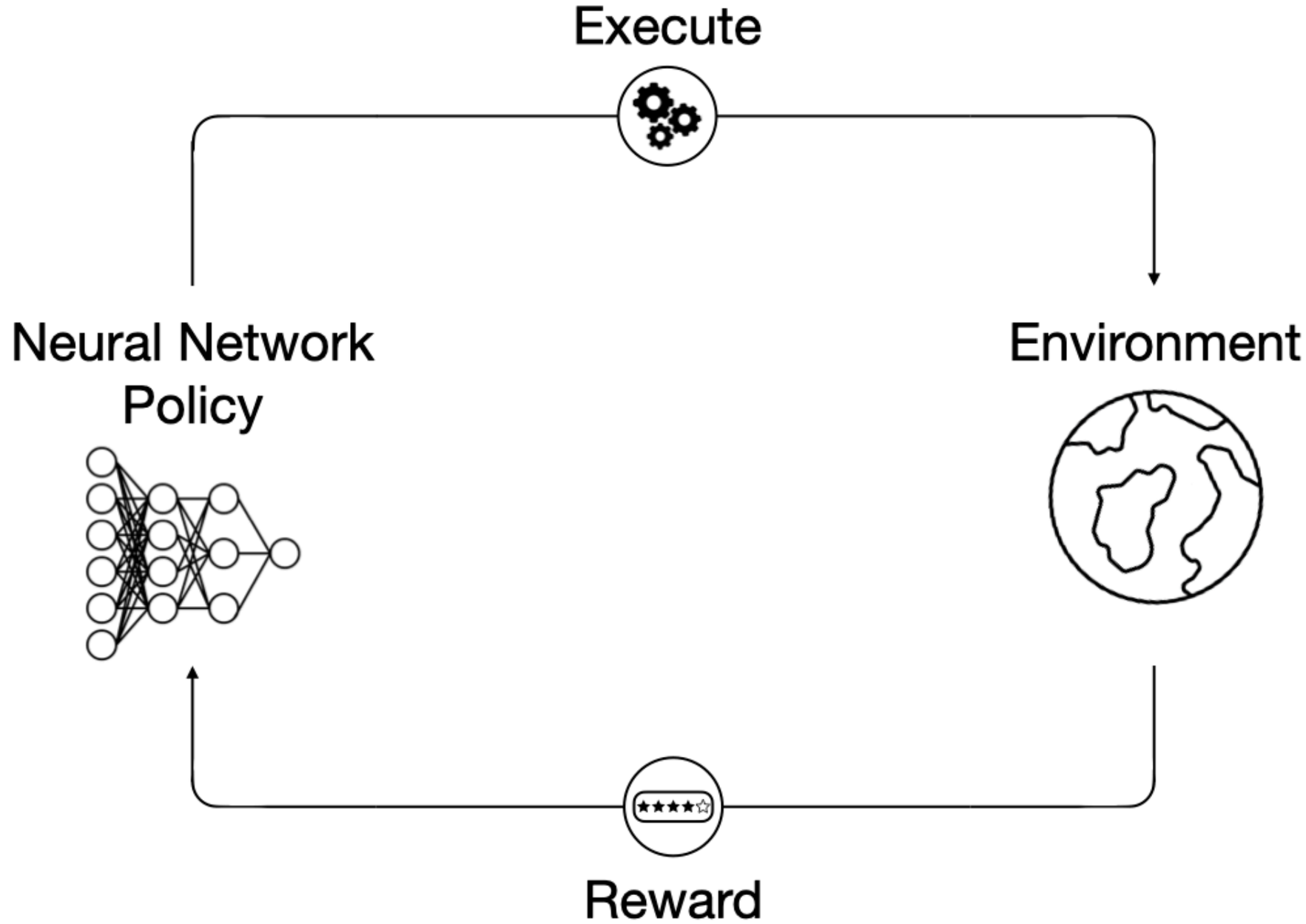


Dweep Trivedi

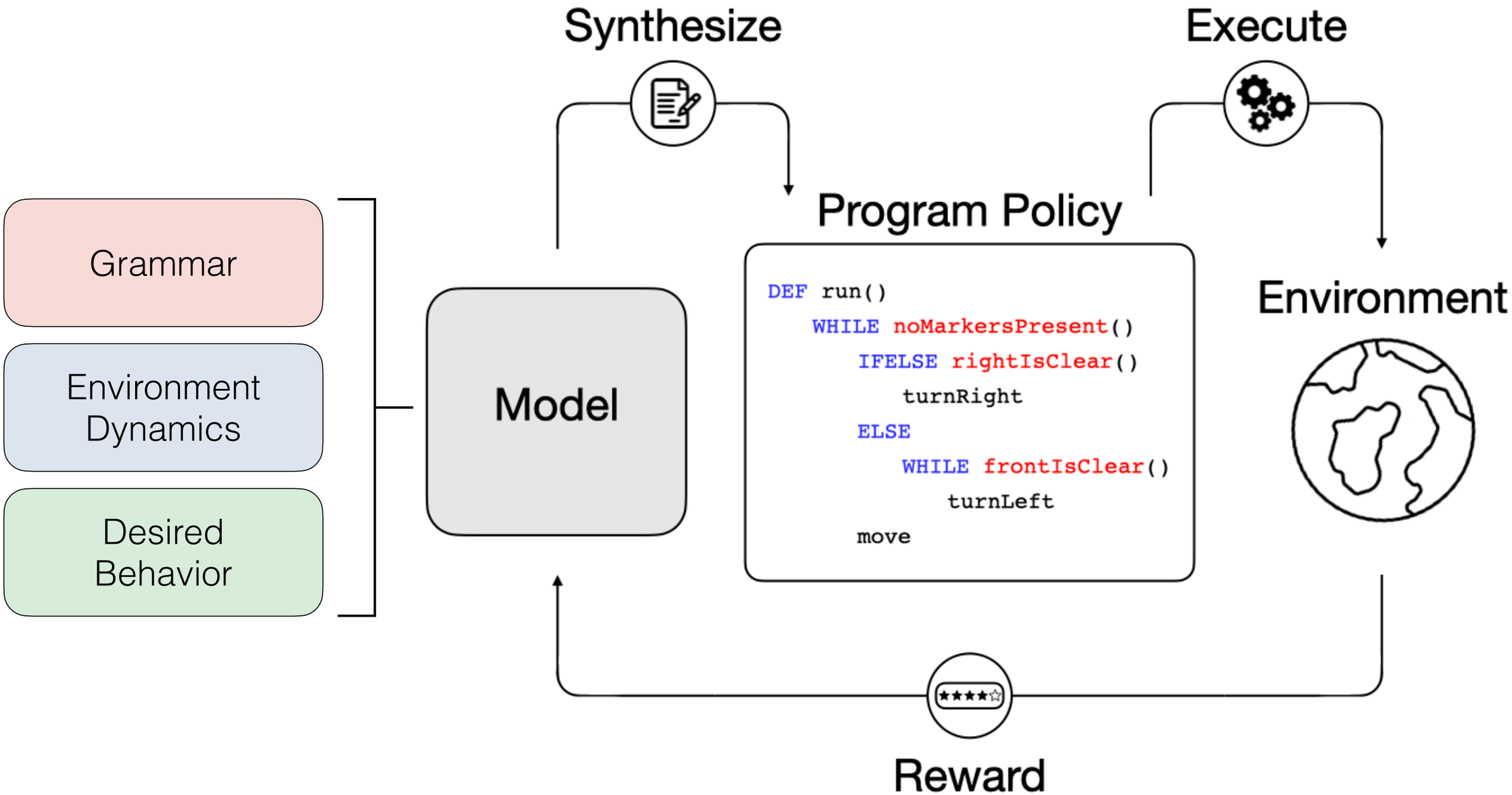

Jesse Zhang


Joseph J. Lim

Reinforcement Learning



Reinforcement Learning by Synthesizing Programs



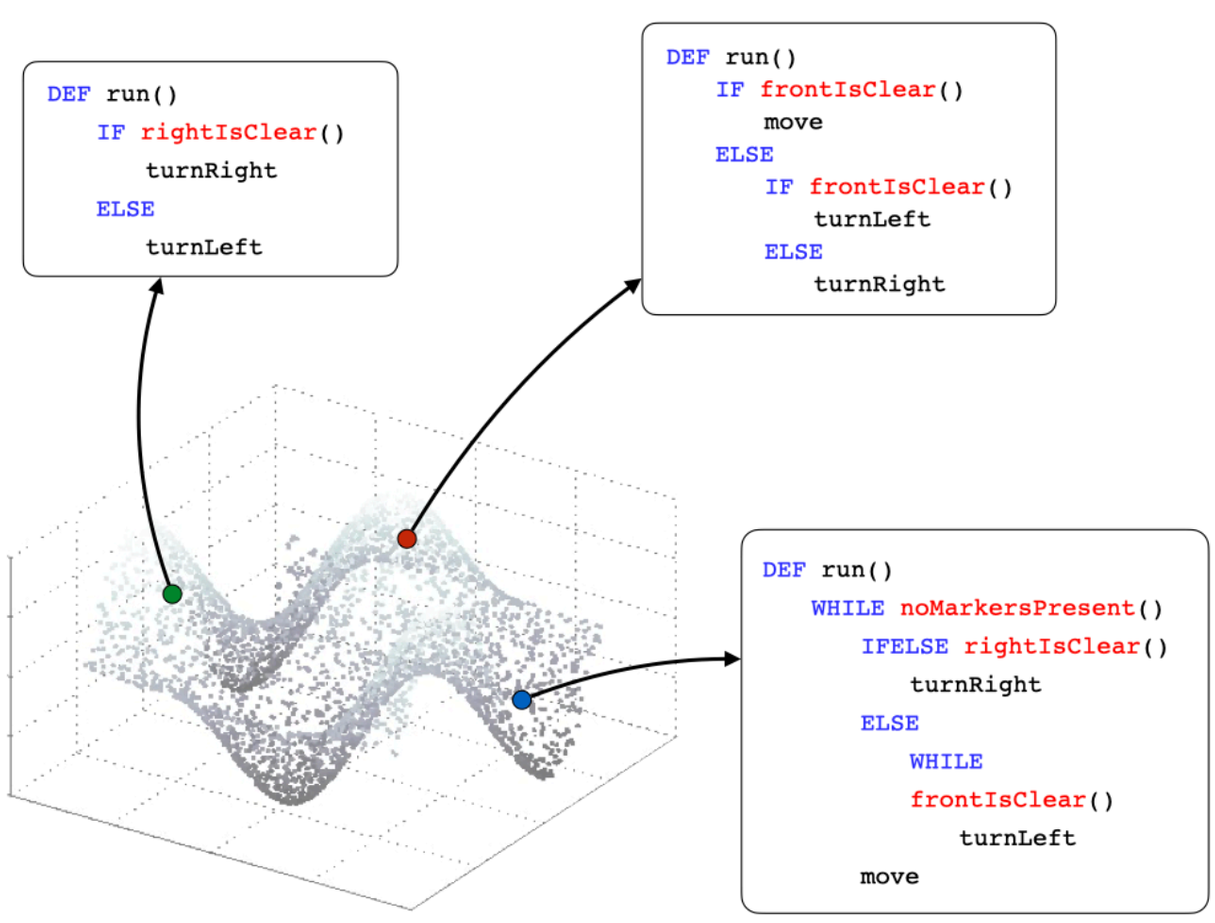
LEAPS: **L**earning **E**mbeddings for l**A**tent **P**rogram **S**ynthesis

Stage 1

Learn a program embedding space from randomly generated programs

Grammar

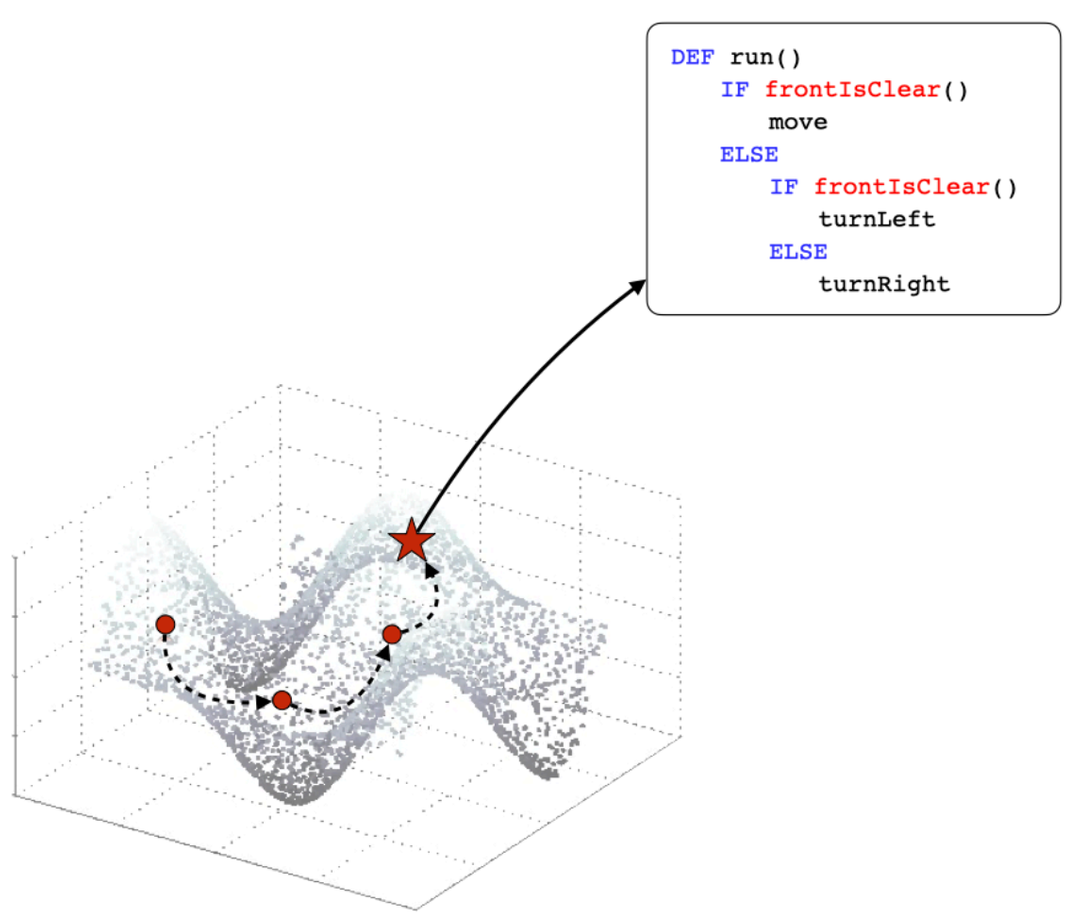
Environment Dynamics



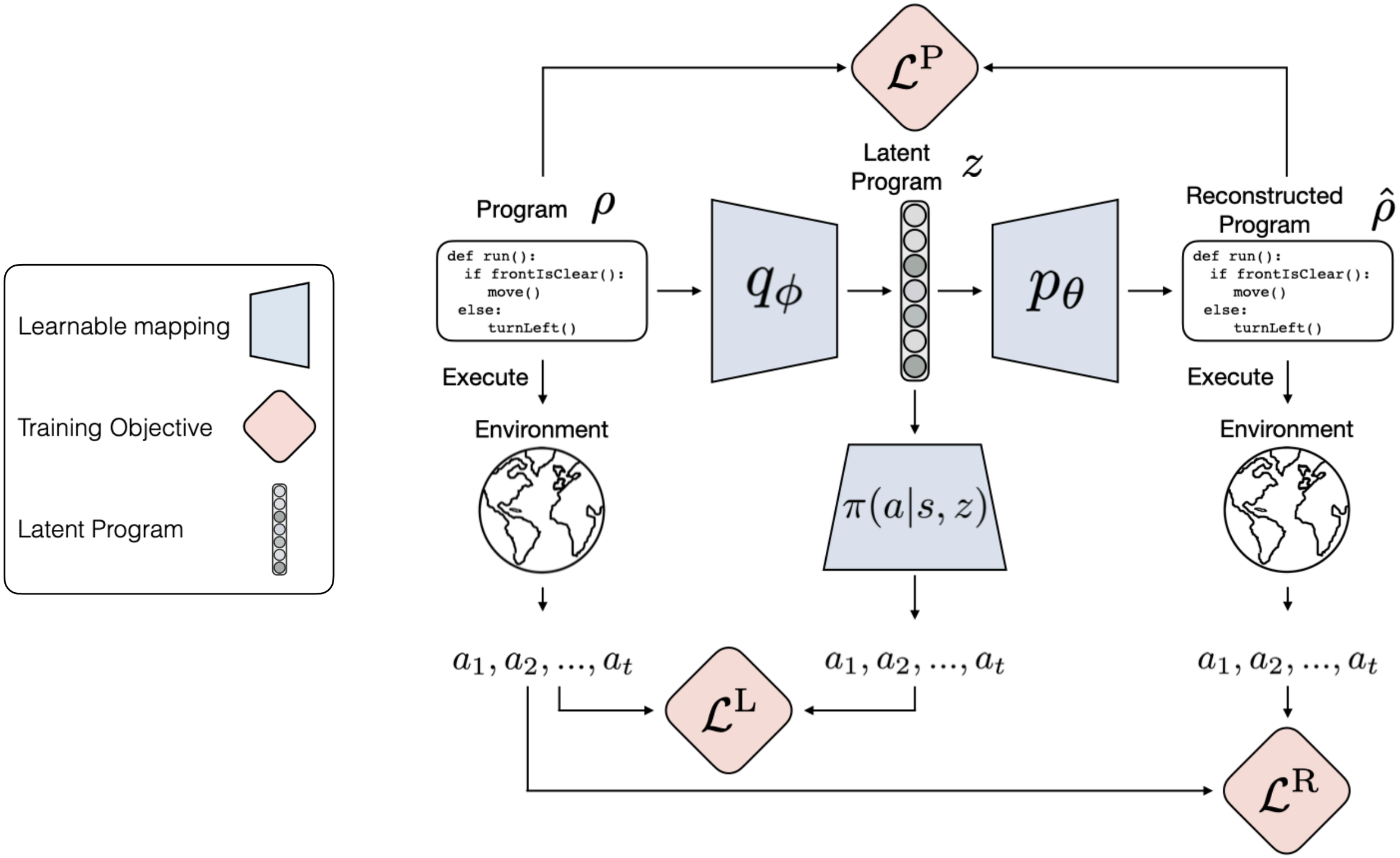
Stage 2

Search for a task-solving program

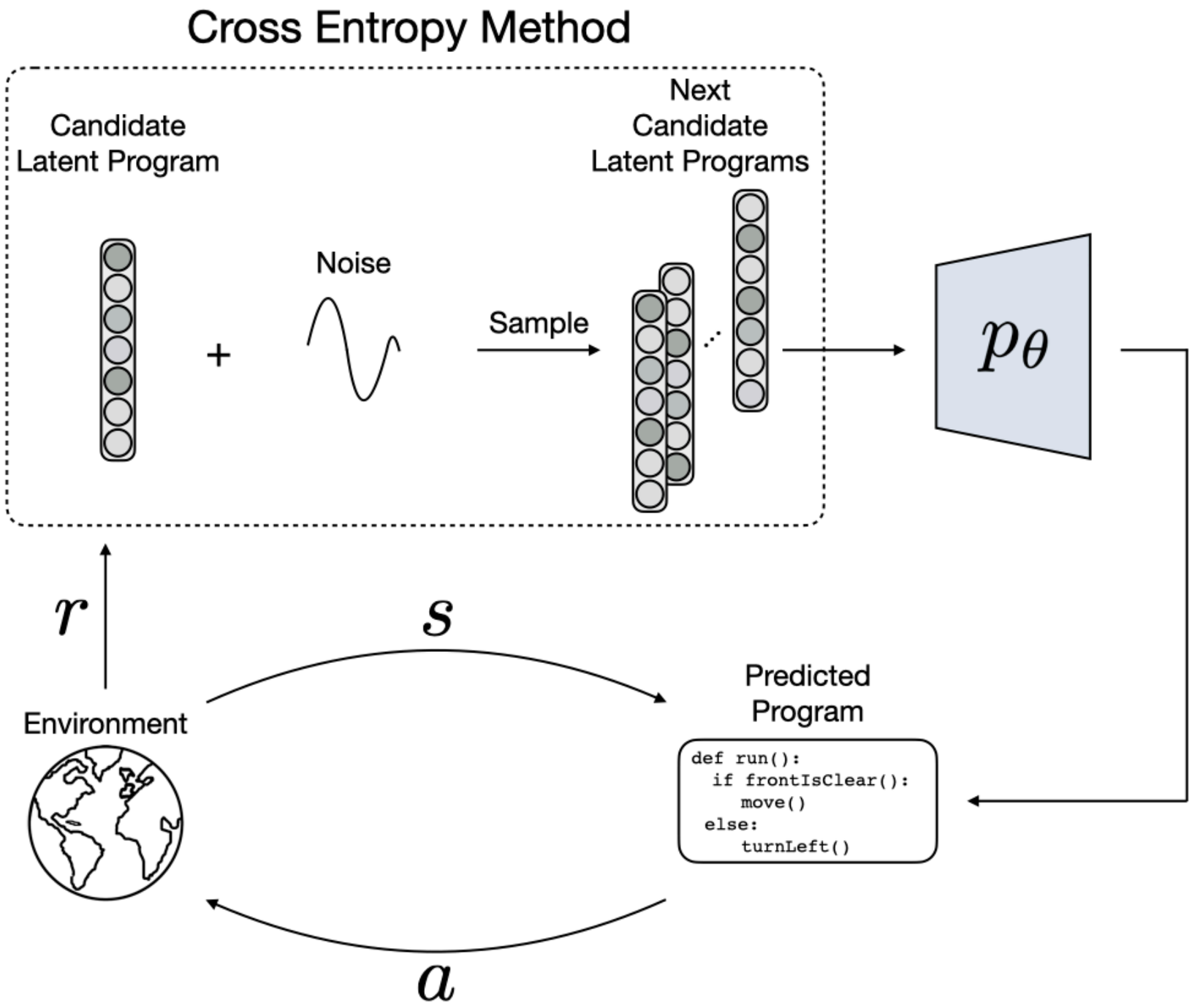
Desired Behavior



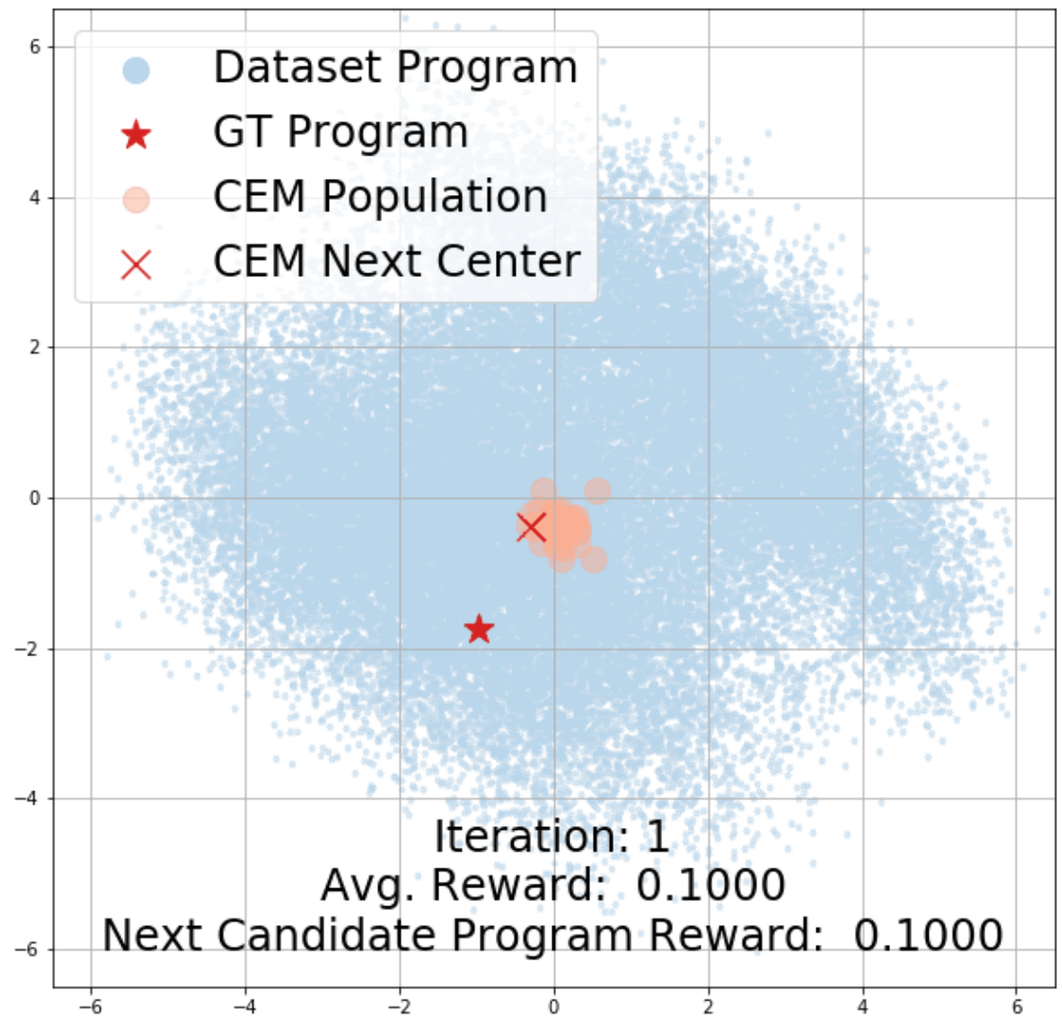
Stage 1: Learning a Program Embedding Space



Stage 2: Latent Program Search with Cross Entropy Method

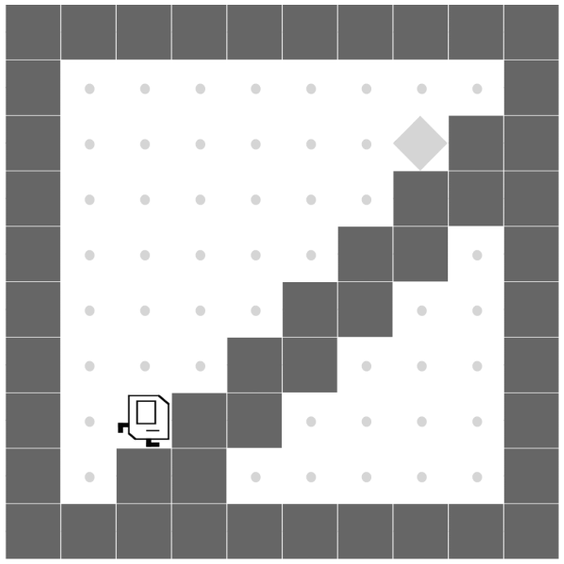


Cross Entropy Method Trajectory Visualization

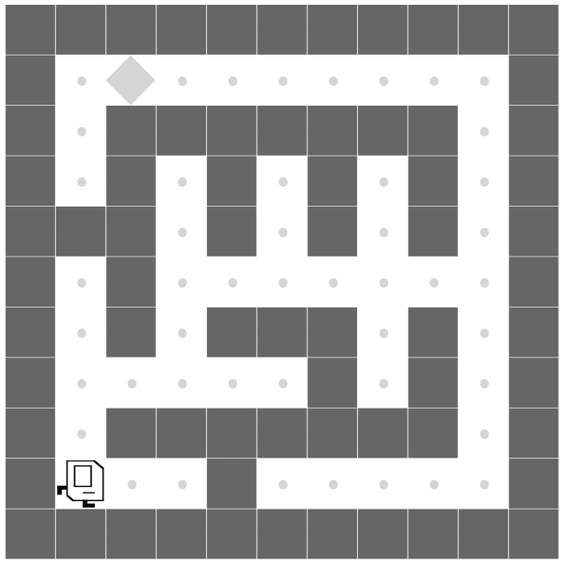


Karel Tasks

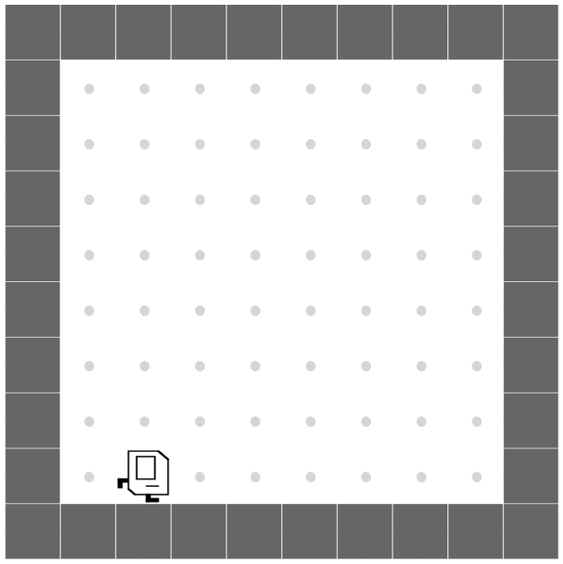
StairClimber



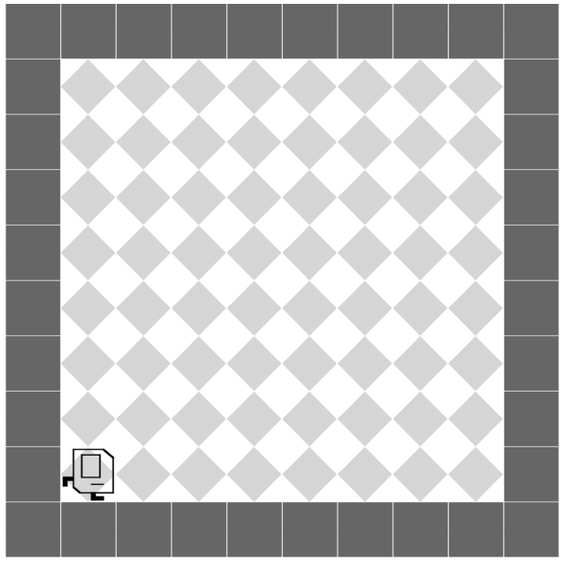
Maze



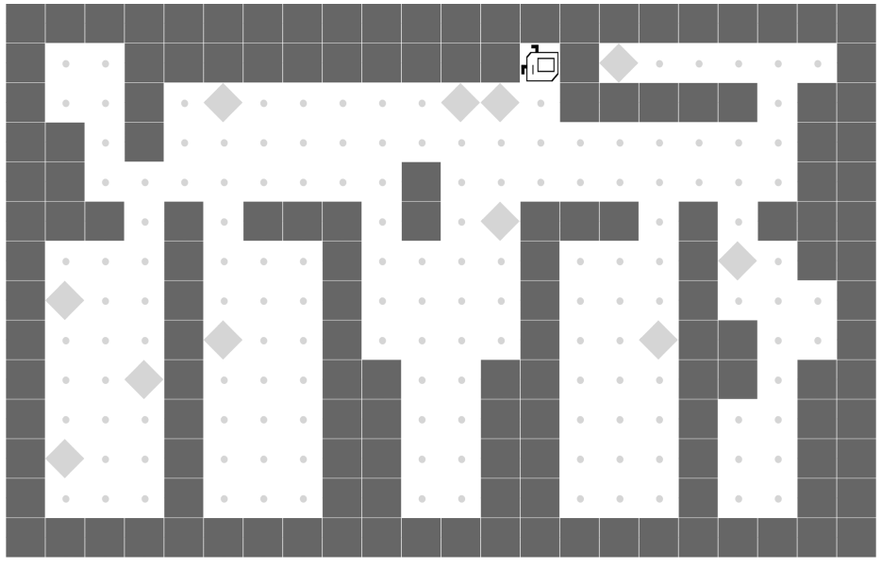
FourCorners



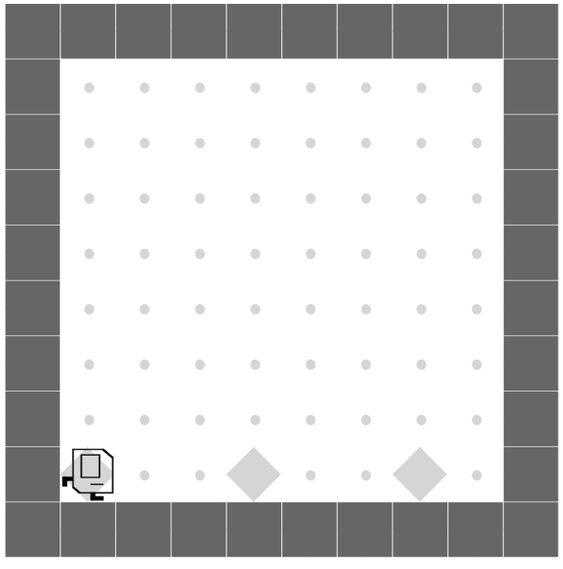
Harvester



CleanHouse

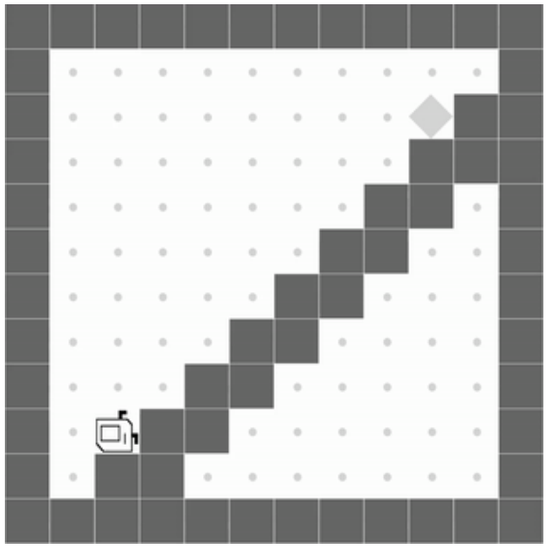


TopOff

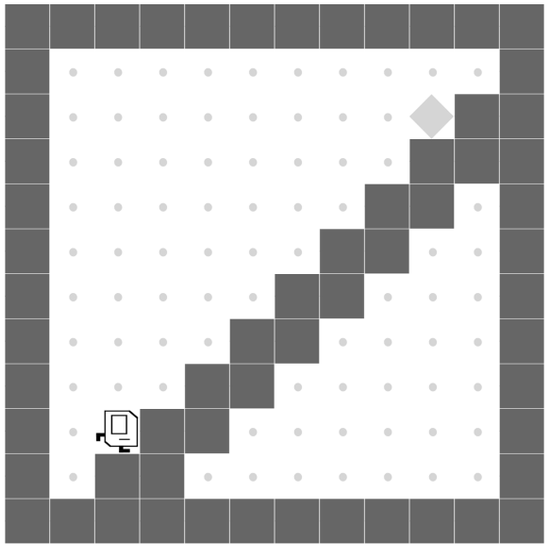


Qualitative Results

StairClimber

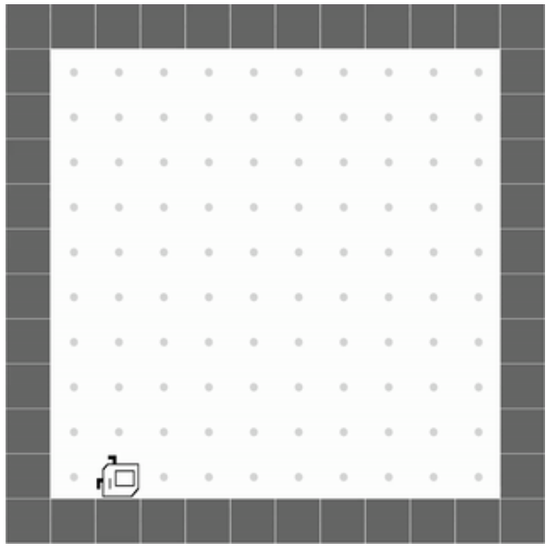


DRL

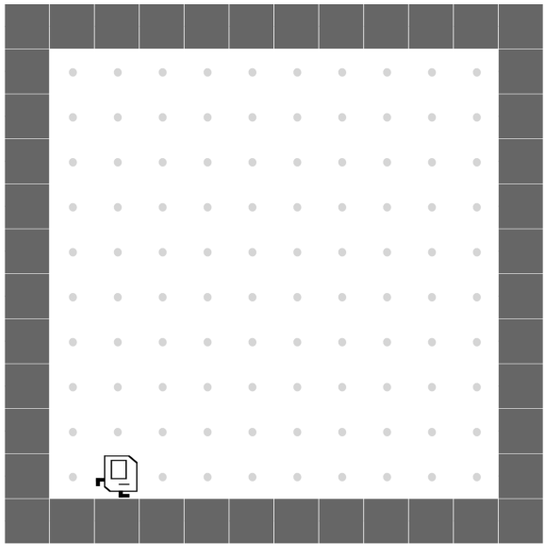


LEAPS

FourCorners

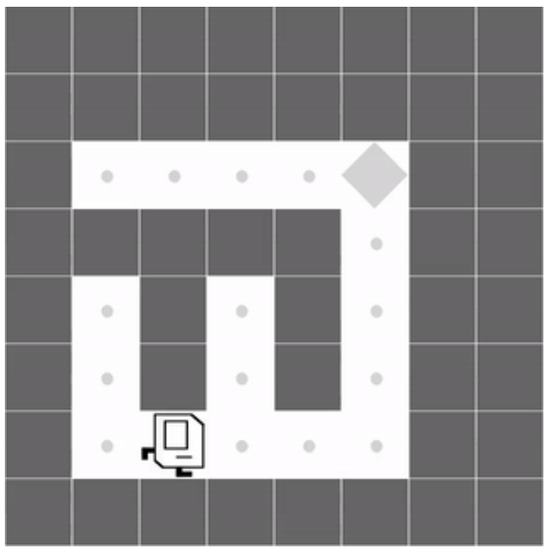


DRL

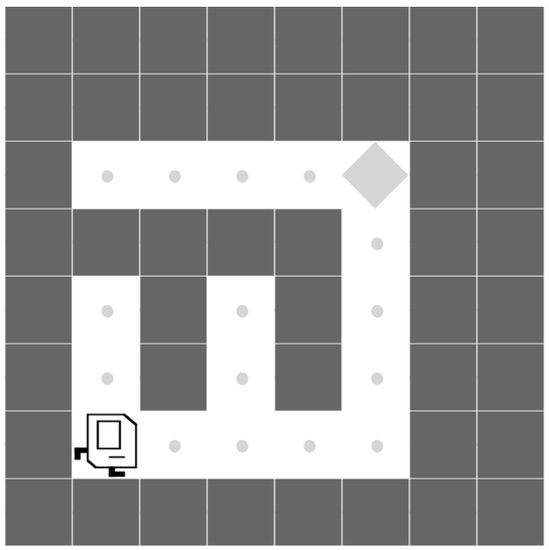


LEAPS

Maze

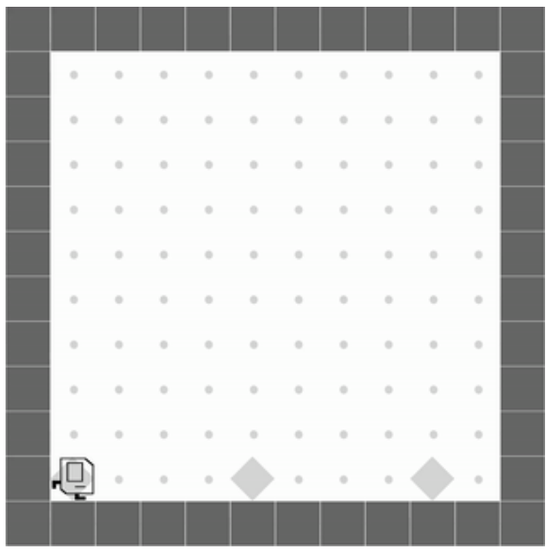


DRL

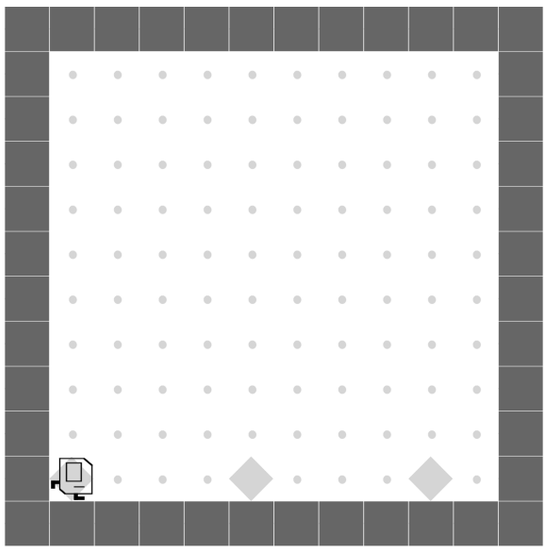


LEAPS

TopOff

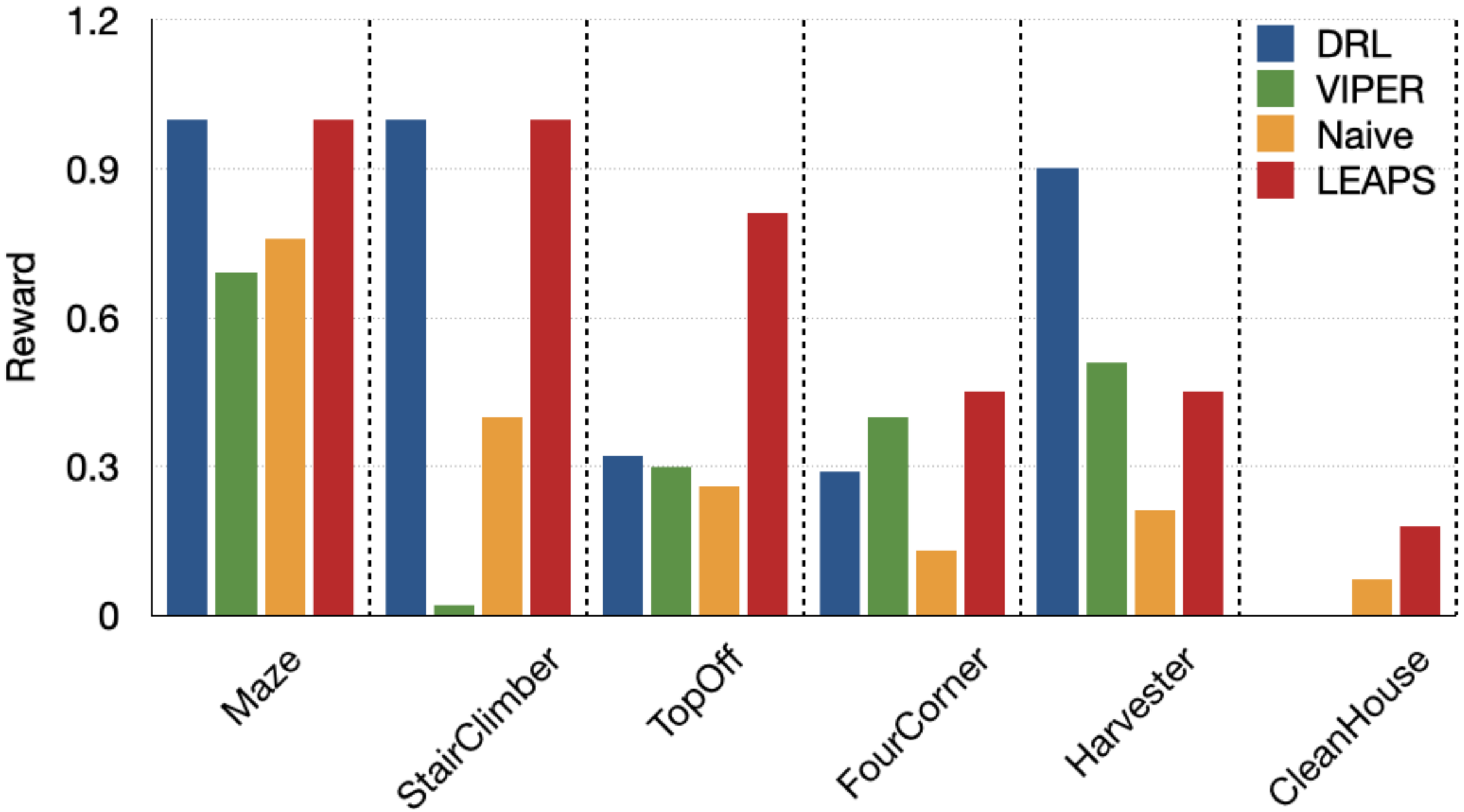


DRL



LEAPS

Quantitative Results

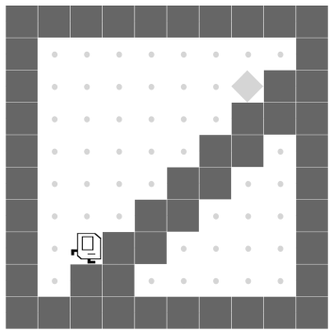


Zero-shot Generalization

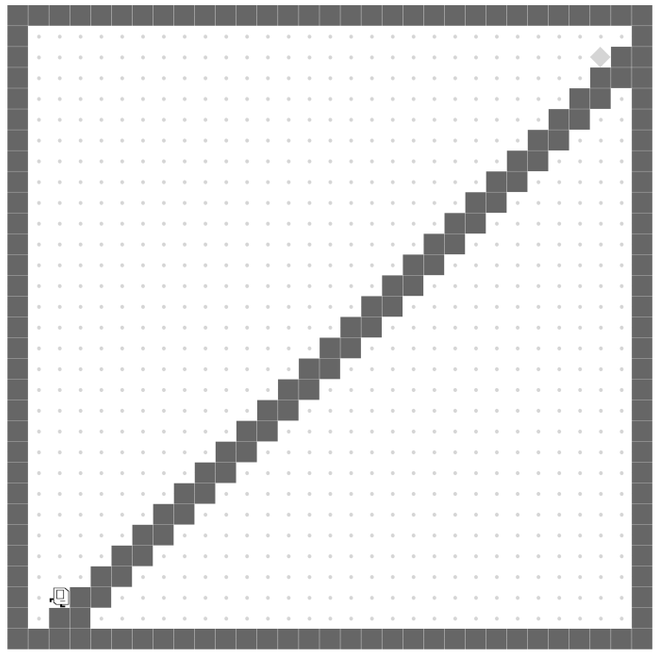
Learning on 8x8 grids

Evaluation on 100x100 grids

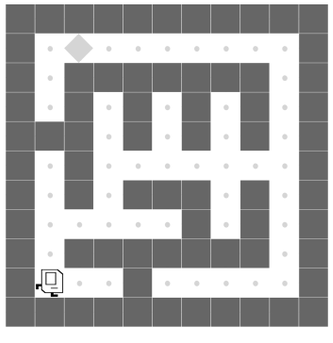
StairClimber



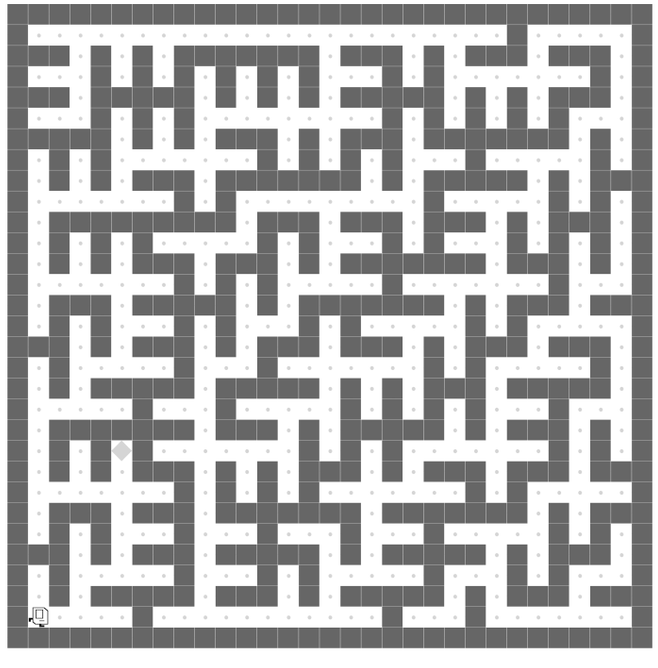
```
DEF run()
   WHILE noMarkersPresent()
      turnRight
      move
   WHILE rightIsClear()
      turnLeft
```



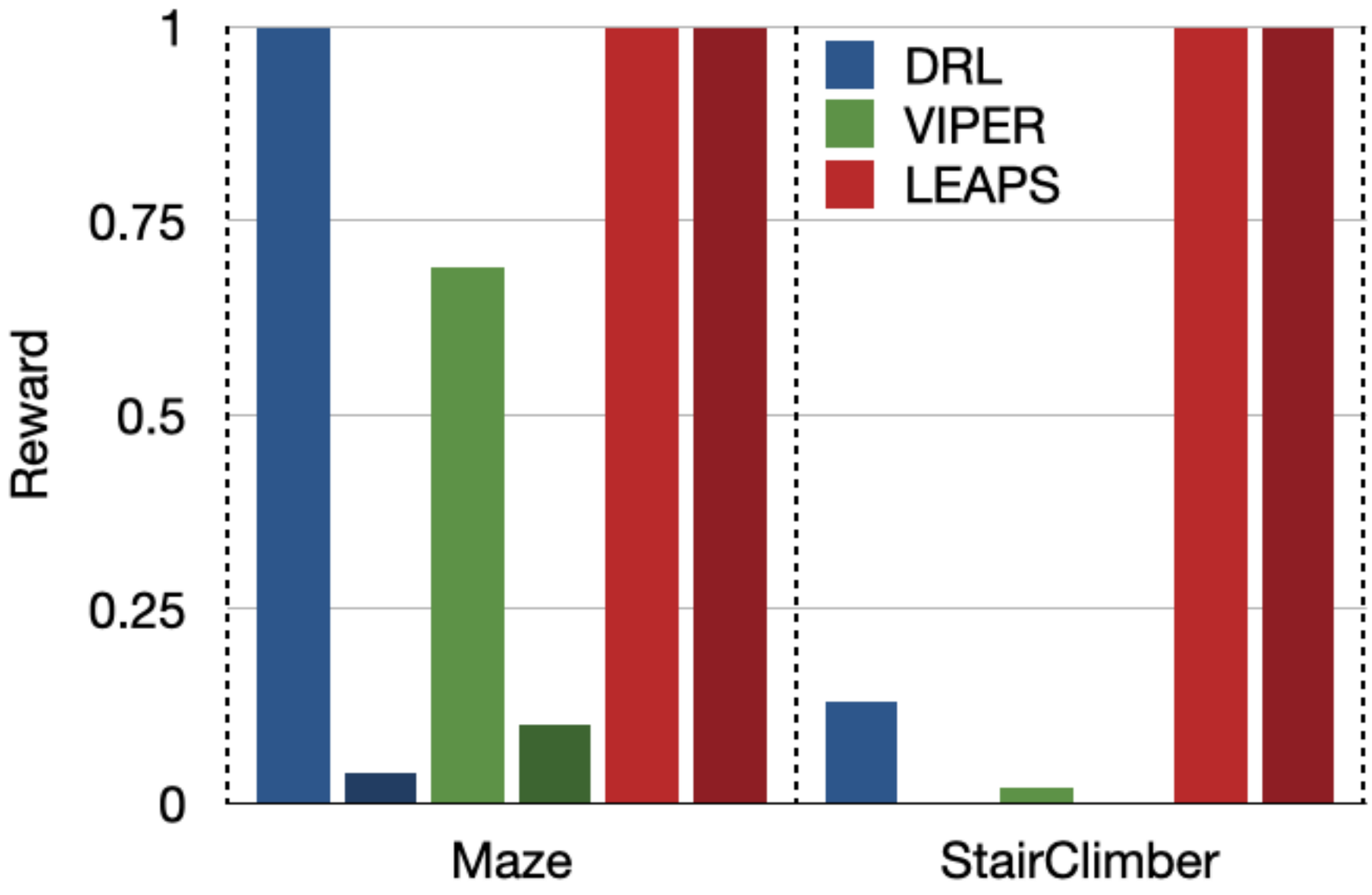
Maze



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

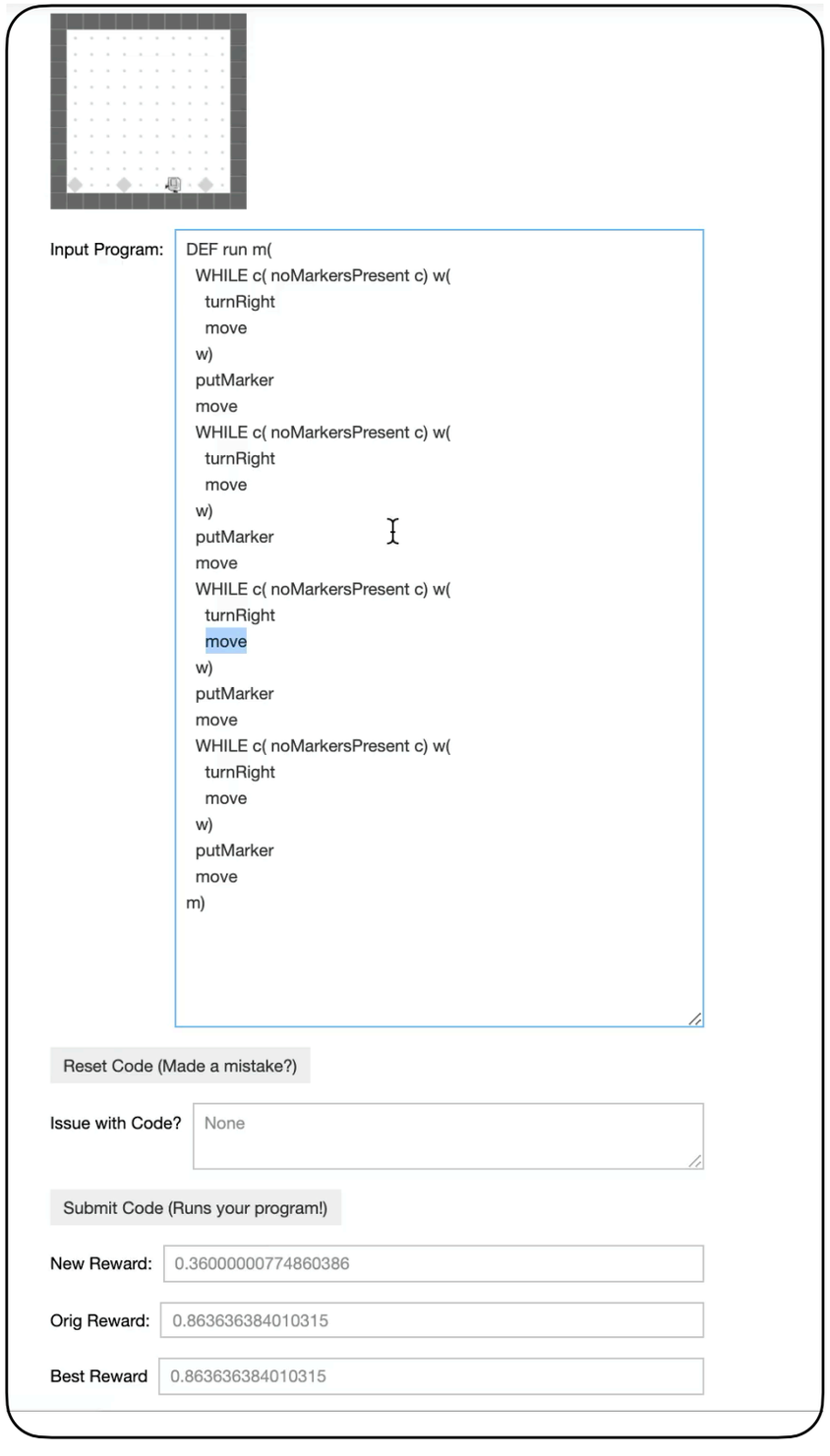


Zero-shot Generalization

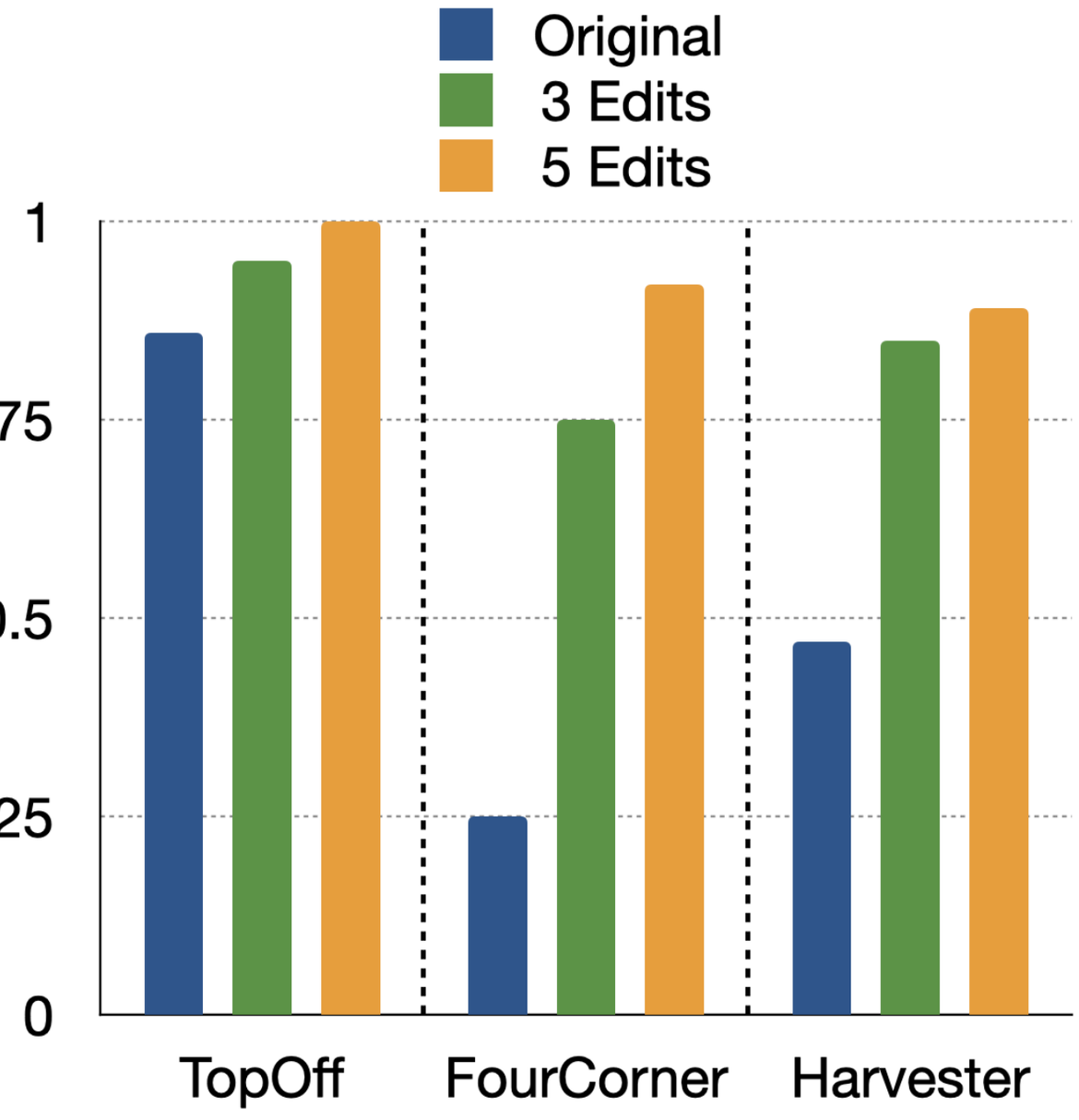


Interpretability

Human Debugging Interface

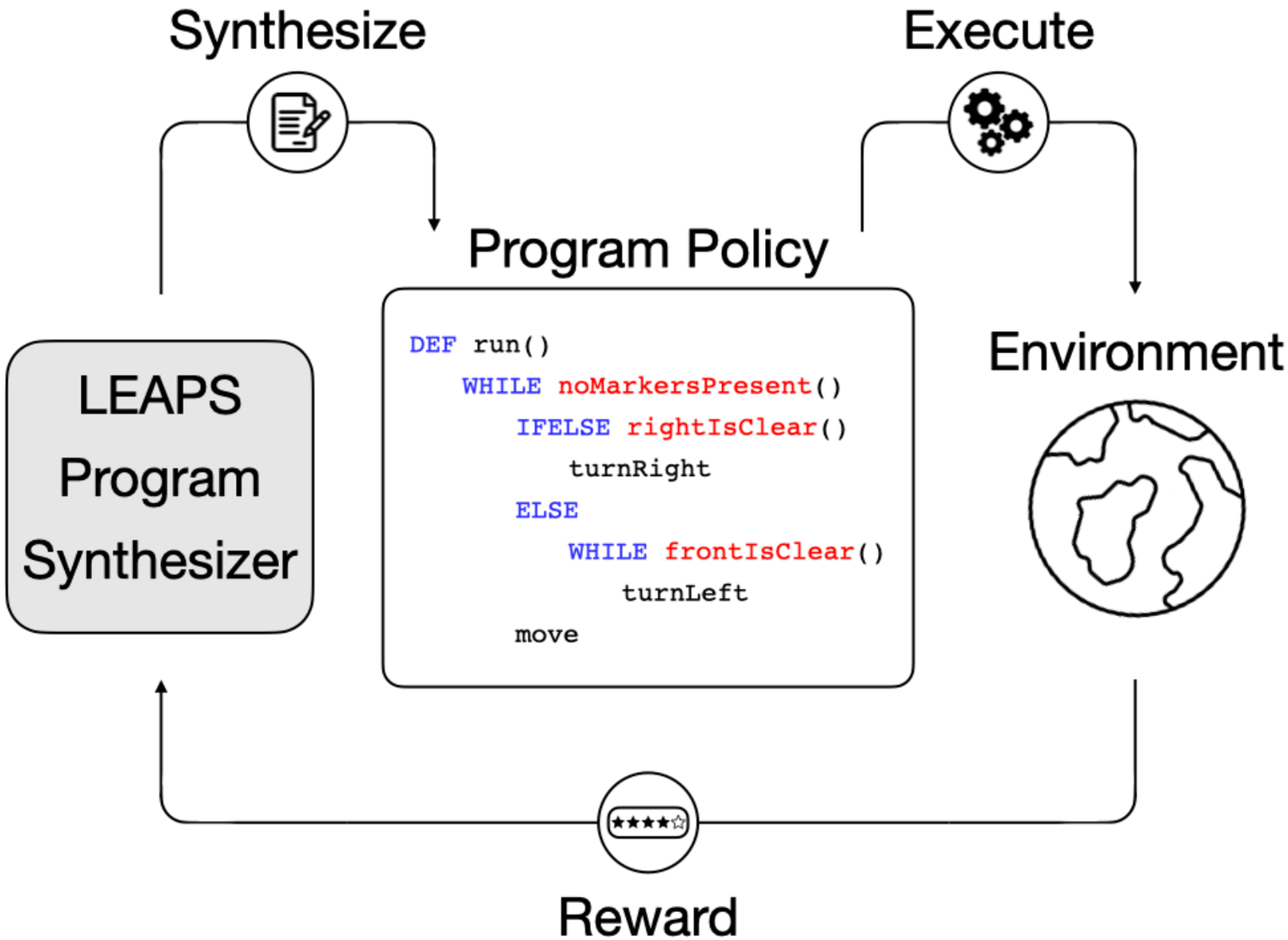


Improved Performance



Takeaway

- Synthesize **generalizable** and **interpretable programs** from rewards

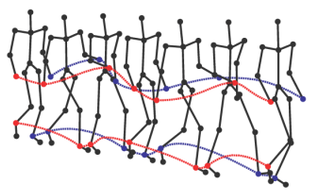


Program Inference

Skill Specification → Program → High-level Plan → Low-level Execution

Video demonstration

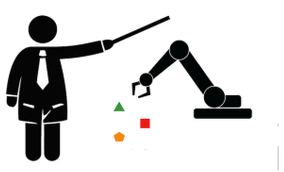
Trajectory demonstration



Language instruction

Stir-fry the onions until tender, and repeat this for garlic and carrots, if you have soy sauce, add some. Pour 2/3 cups the whisked eggs into the stir-fried and scramble.

Reward



Actions (sub-skills)

`Jump` `Crawl` `Walk`

Perceptions

`isCurb()` `isCeiling()`

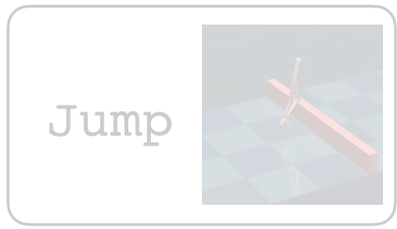
Control Flows

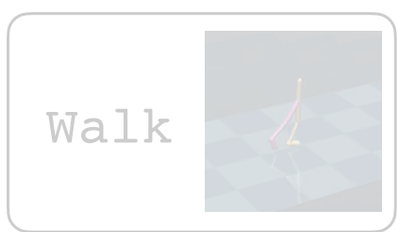
`while` `if` `else` `elseIf`

Program

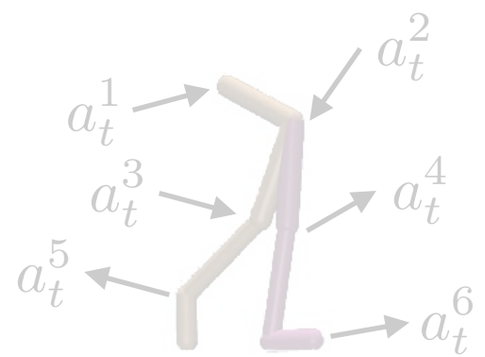
```
DEF run()
    while True:
        if isCurb():
            jump
        elseIf isCeiling():
            crawl()
        else:
            walk()
```

Observation



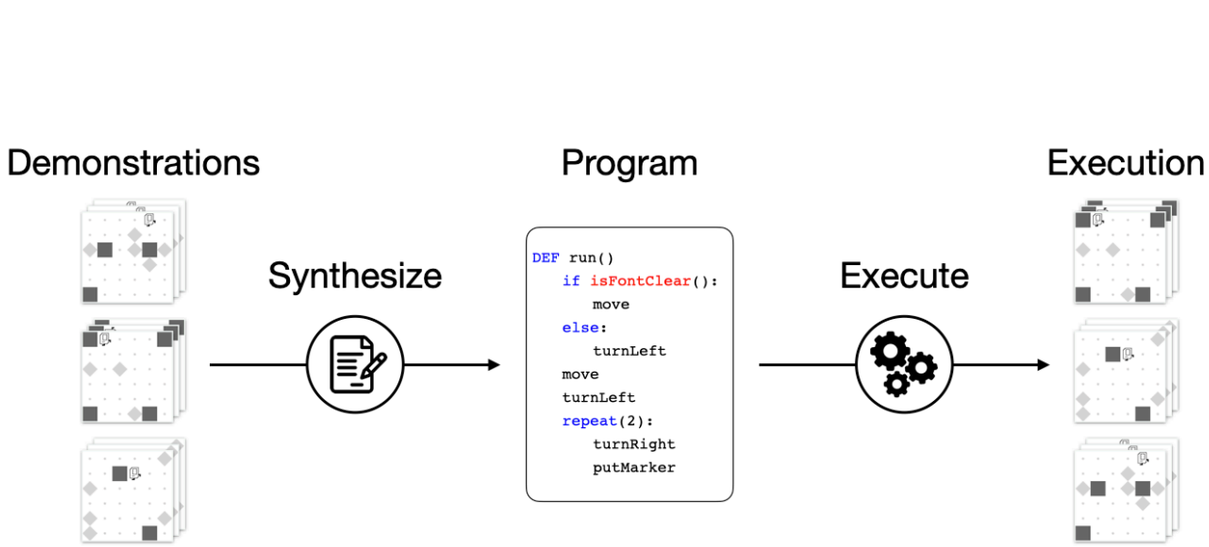


Low-level execution

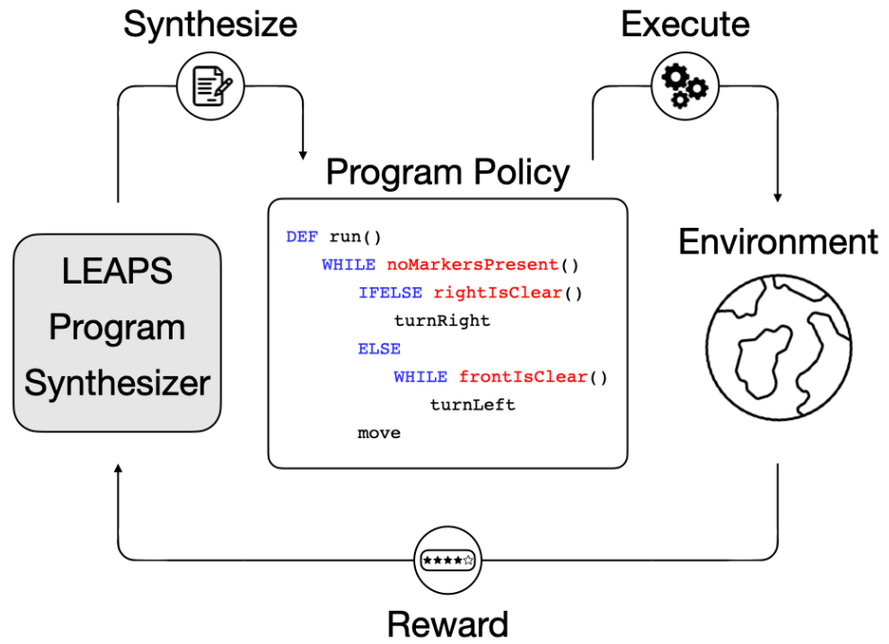


Program Inference

Imitation learning from demonstrations

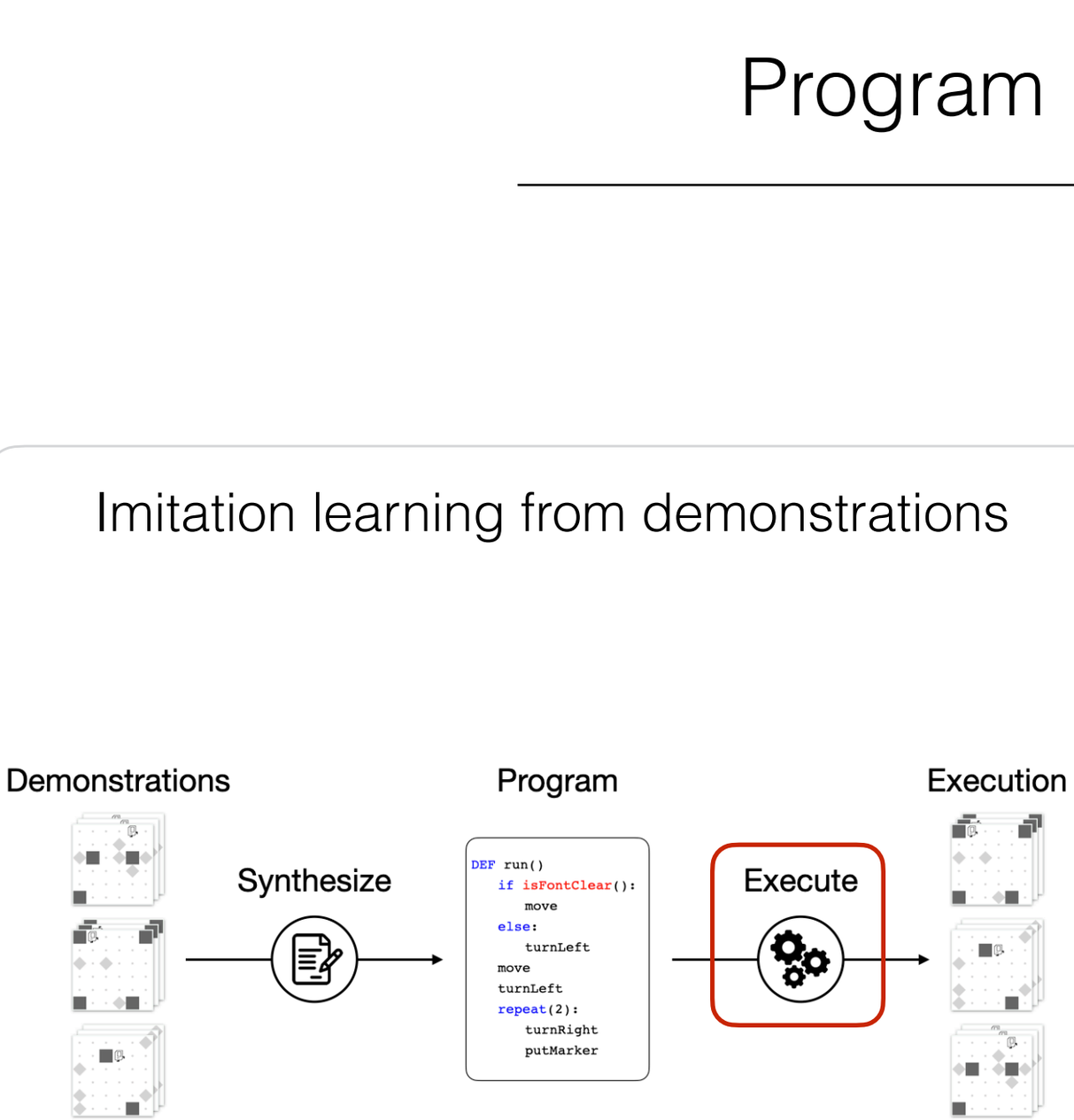


Reinforcement learning from rewards

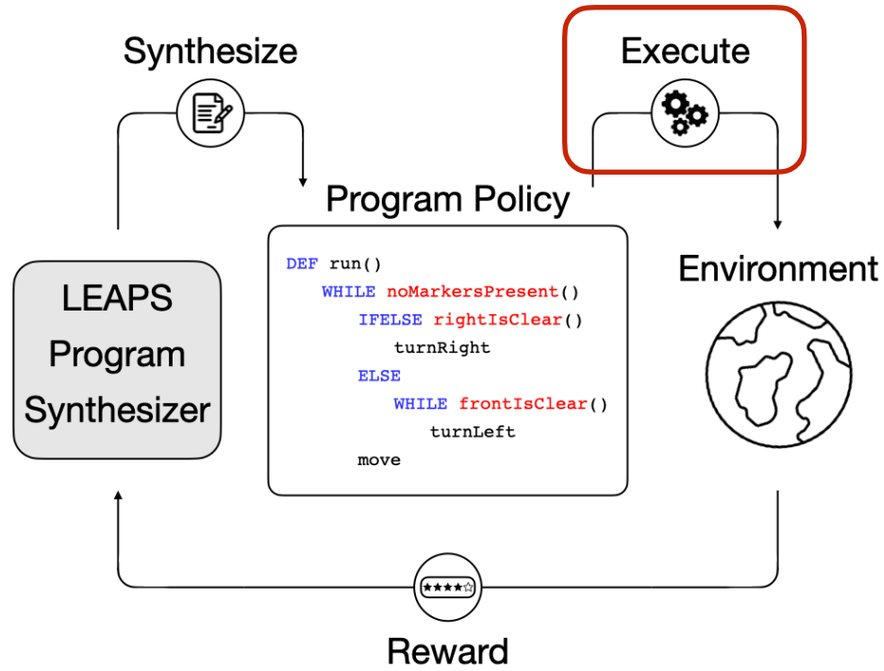


Program Inference

Imitation learning from demonstrations



Reinforcement learning from rewards



Task Execution

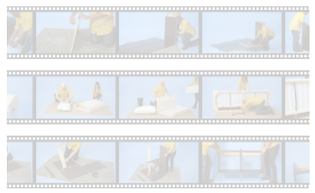
Skill Specification

Program

High-level Plan

Low-level Execution

Video demonstration



Trajectory demonstration

Language instruction

Stir-fry the onions until tender, and repeat this for garlic and carrots, if you have soy sauce, add some. Pour 2/3 cups the whisked eggs into the stir-fried and scramble.

Reward

Actions (sub-skills)

`Jump` `Crawl` `Walk`

Perceptions

`isCurb()` `isCeiling()`

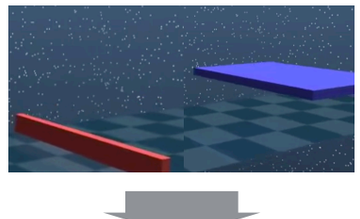
Control Flows

`while` `if` `else` `elseIf`

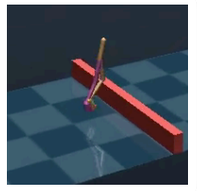
Program

```
DEF run()
    while True:
        if isCurb():
            jump
        elseIf isCeiling():
            crawl()
        else:
            walk()
```

Observation



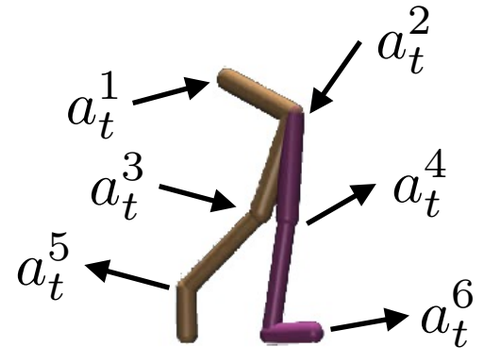
`Jump`



`Walk`

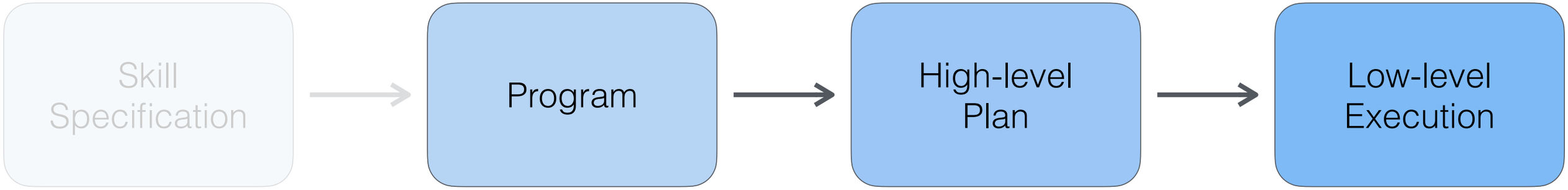
`Crawl`

Low-level execution



Program Guided Agent

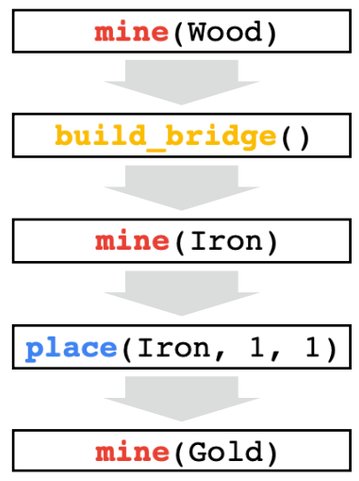
ICLR 2020 (Spotlight)

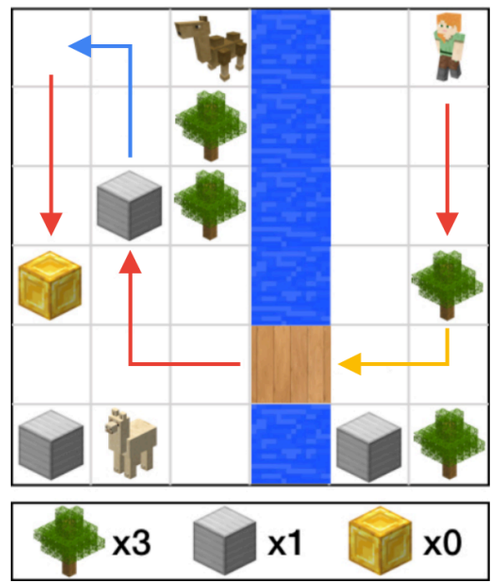



Te-Lin Wu


Joseph J. Lim


```
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)
```



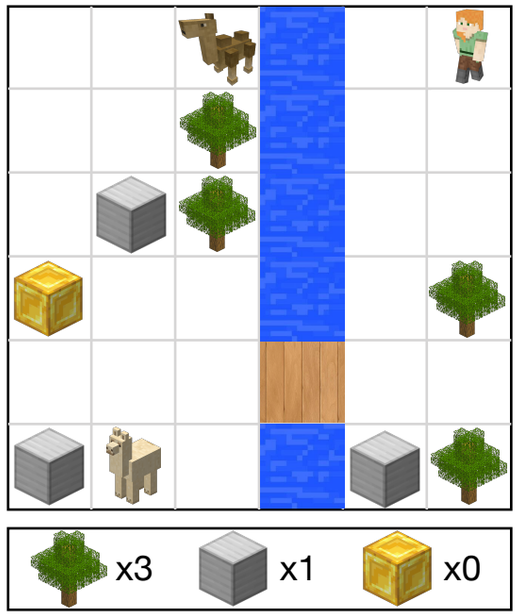


Problem Formulation

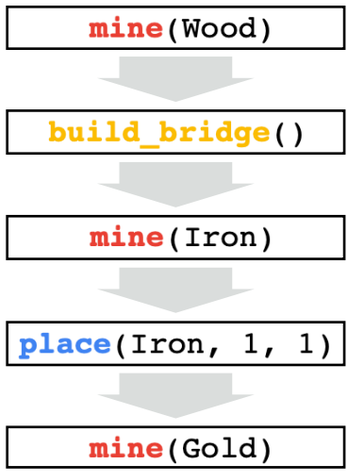
Program (task)

```
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)
```

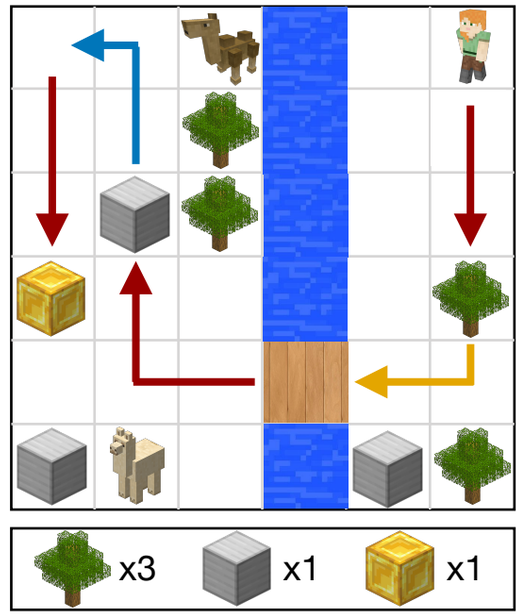
Observation



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)
```

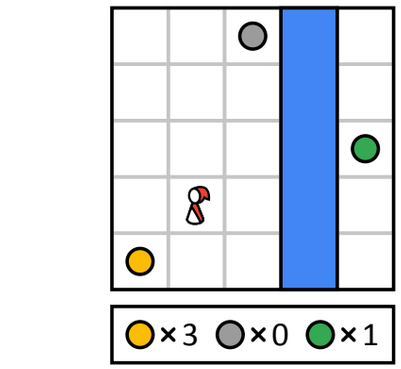
```
def run()
   while agent[Wood] <= 11:
      place(Wood, 2, 4)
      place(Iron, 1, 1)
      place(Iron, 8, 5)
      mine(Gold)
      mine(Gold)
   mine(Gold)
   repeat(4):
      sell(Gold)
      sell(Iron)
```

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



State

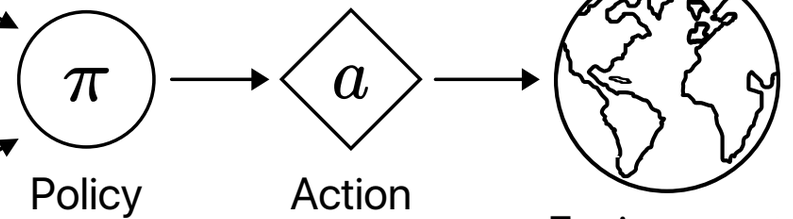
```
def run():
  while env[Gold] > 0:
    mine(Gold)
  if is_there[River]:
    build_bridge()
    place(Wood, 2, 3)
```

Program

```
If an agent has more
than 1 iron, place an
iron on (2,3), and then
if there are less than 3
gold in the environment,
mine gold; otherwise,
goto (4,2). While gold
in the environment is
larger than 2, keep
mining gold.
```

Language description

OR

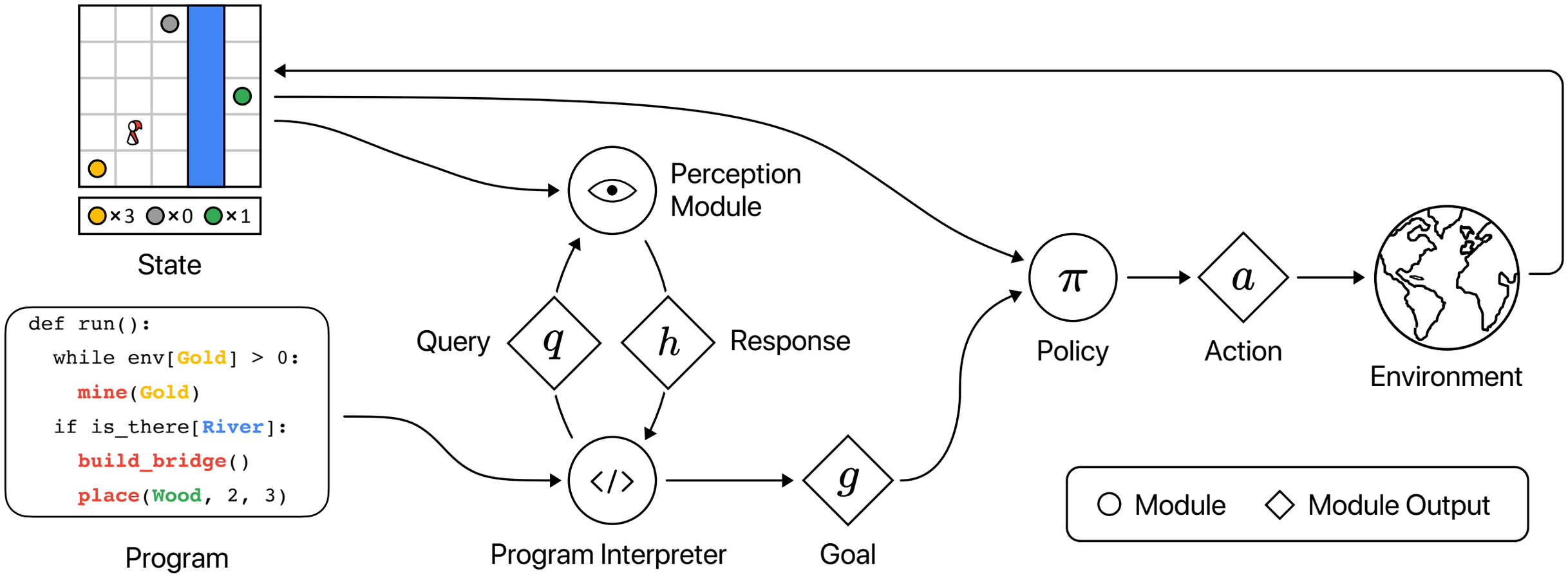


Policy

Action

Environment

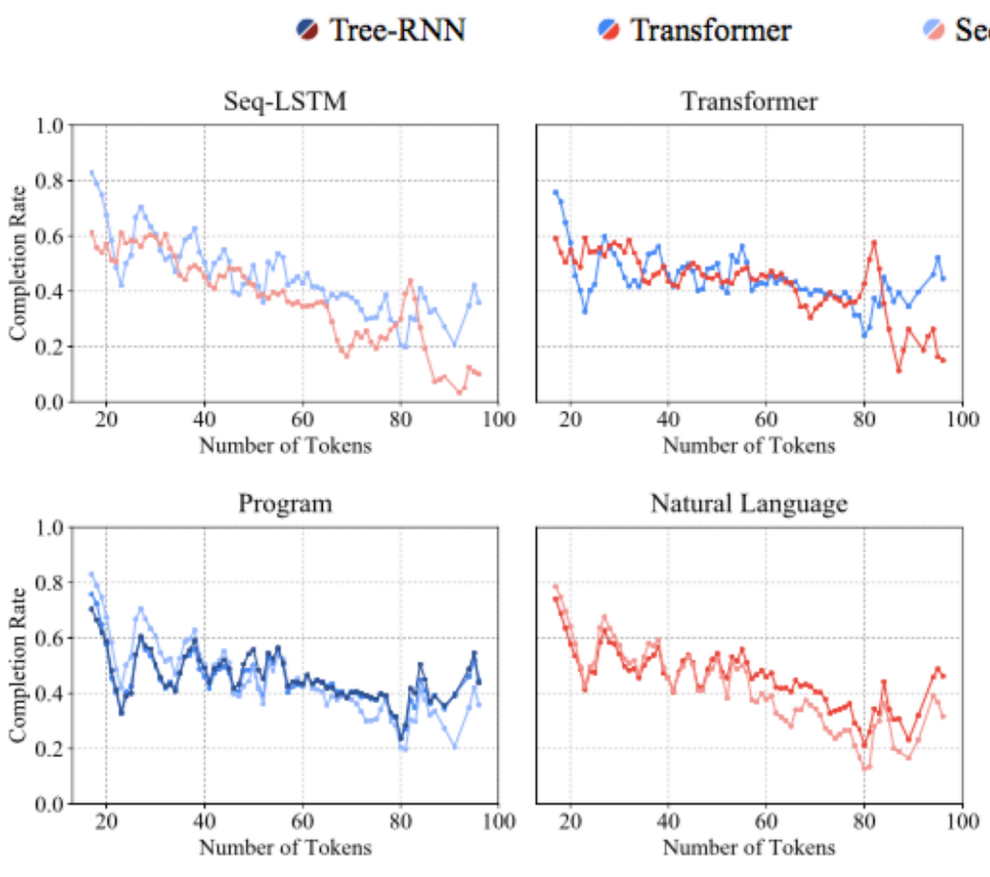
Program Guided Agent



Quantitative Results

Generalization

Instruction		Natural language descriptions		Programs				
Method		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%



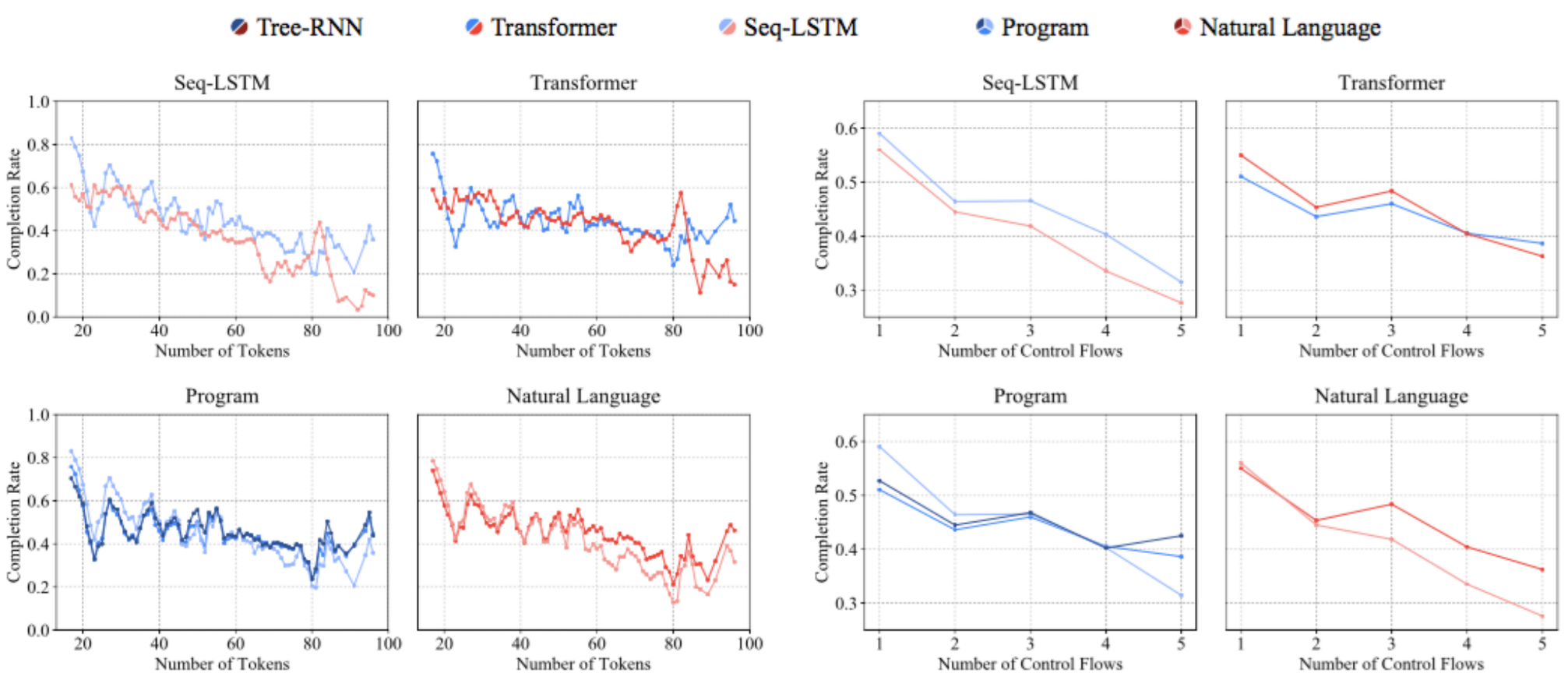
(a) Instruction Length

(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%



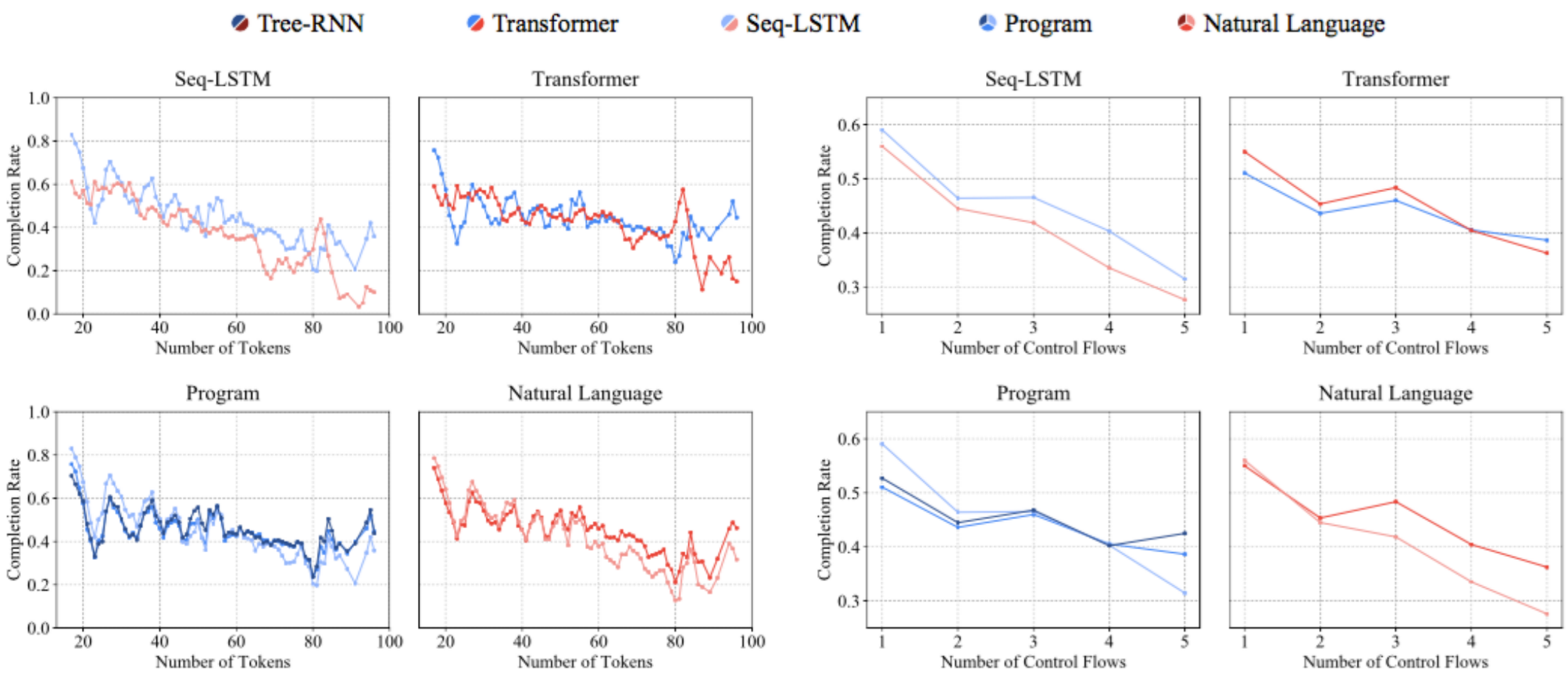
(a) Instruction Length

(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

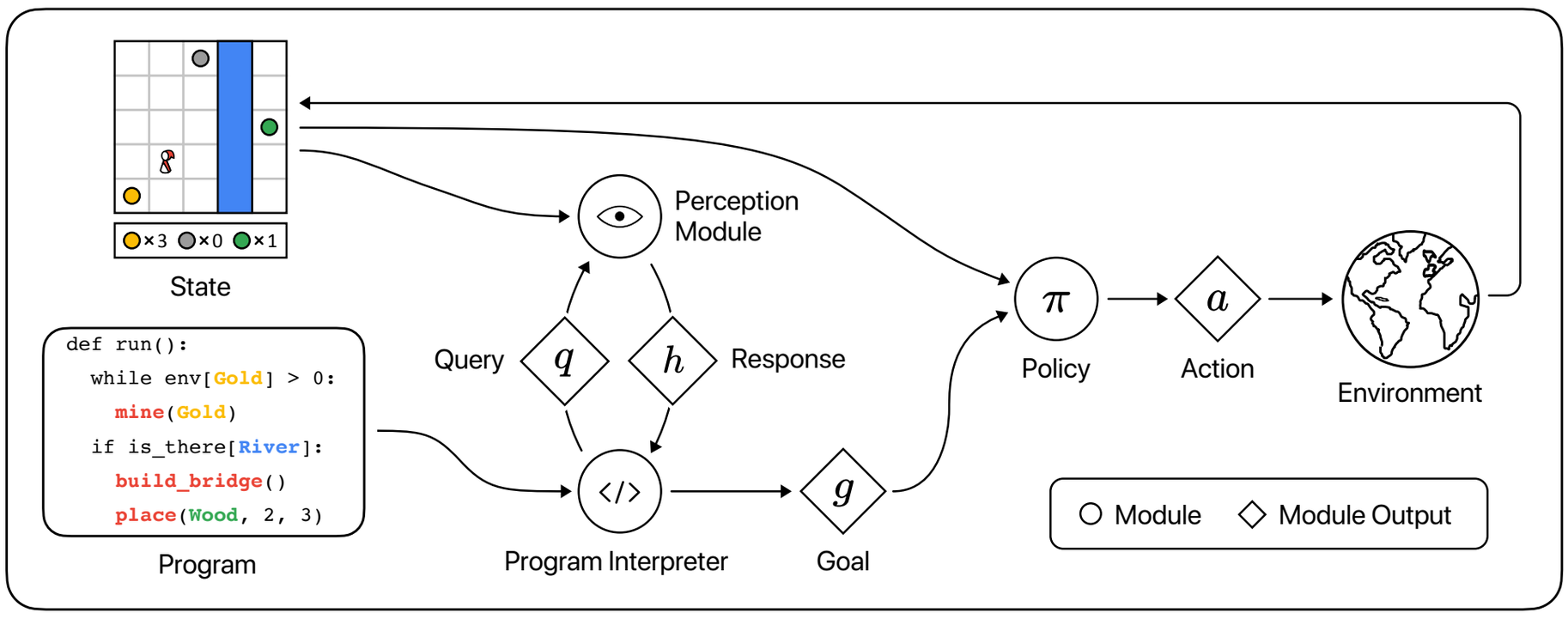
Takeaway

- Specific tasks using **programs**

Program

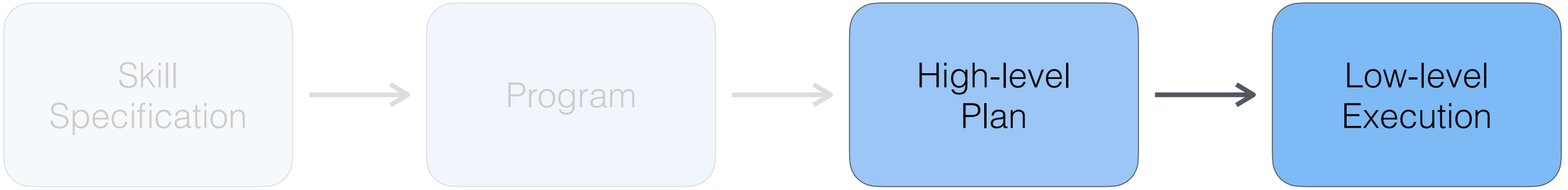
```
def run():
  if is_there[River]:
    mine(Wood)
    build_bridge()
    if agent[Iron]<3:
      mine(Iron)
    place(Iron, 1, 1)
  else:
    goto(4, 2)
  while env[Gold]>0:
    mine(Gold)
```

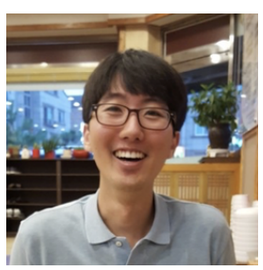
- Leverage the structure of programs with **a modular framework**



Composing Complex Skills by Learning Transition Policies

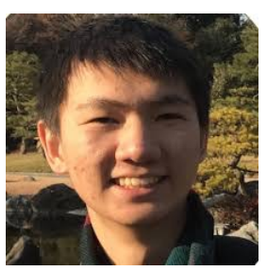
ICLR 2019





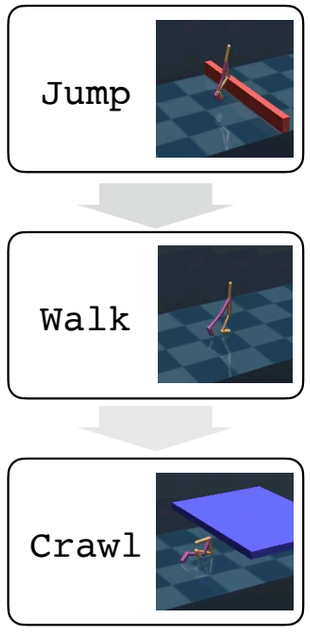
Youngwoon Lee

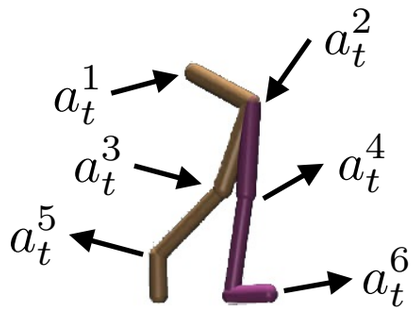
Sriram Somasundaram



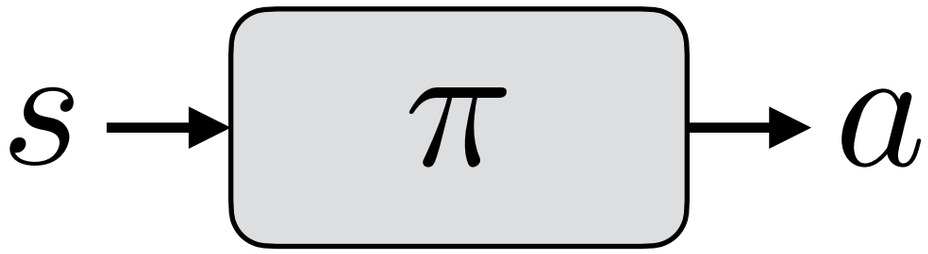
Edward Hu

Joseph J. Lim

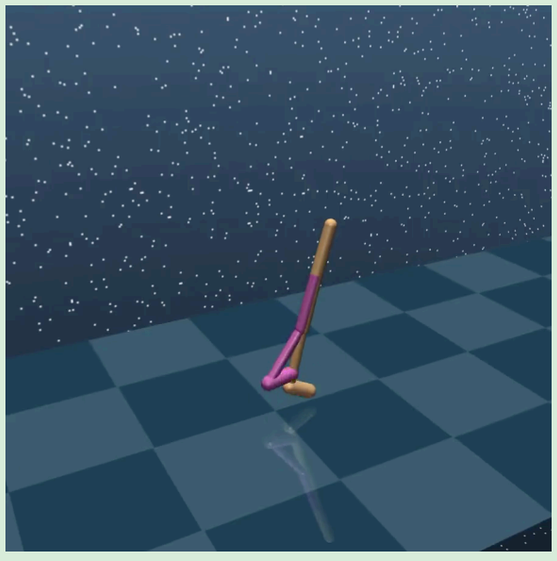




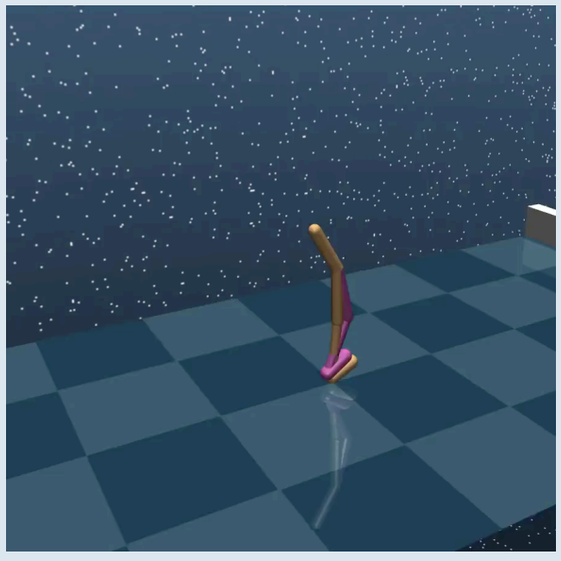
Learned Skills



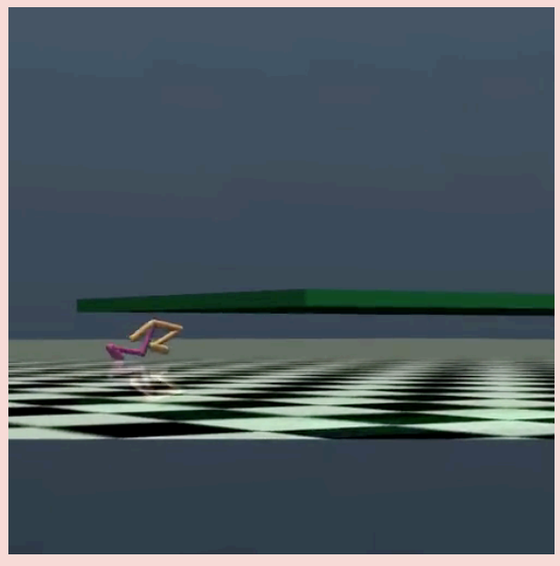
Walk



Jump

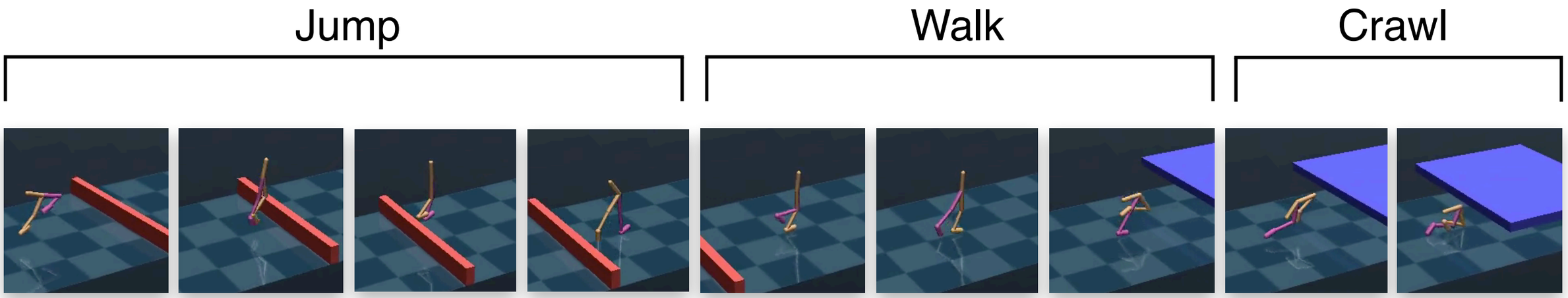


Crawl

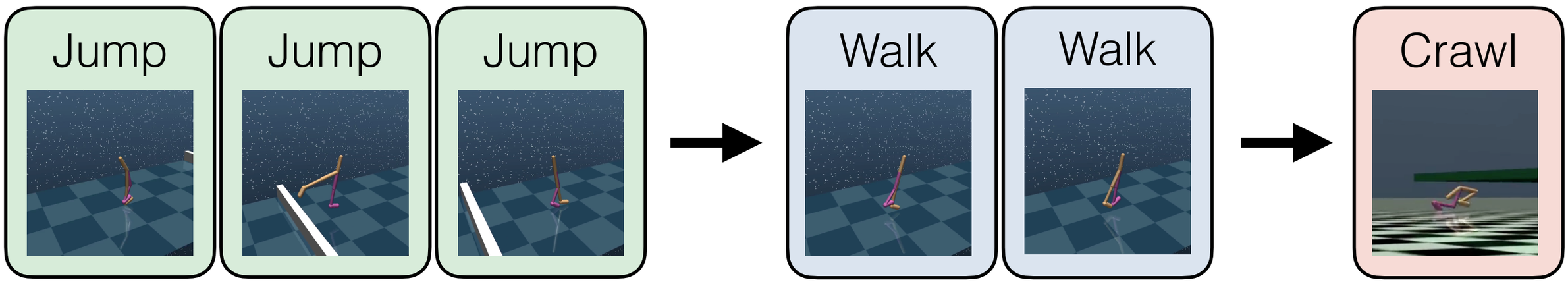


Compose Complex Skills

High-level plan

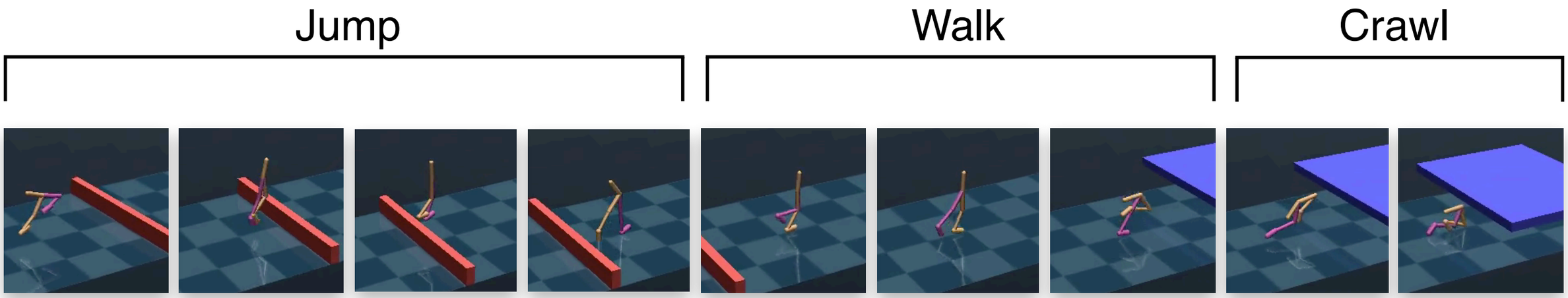


Sequentially execute corresponding policies

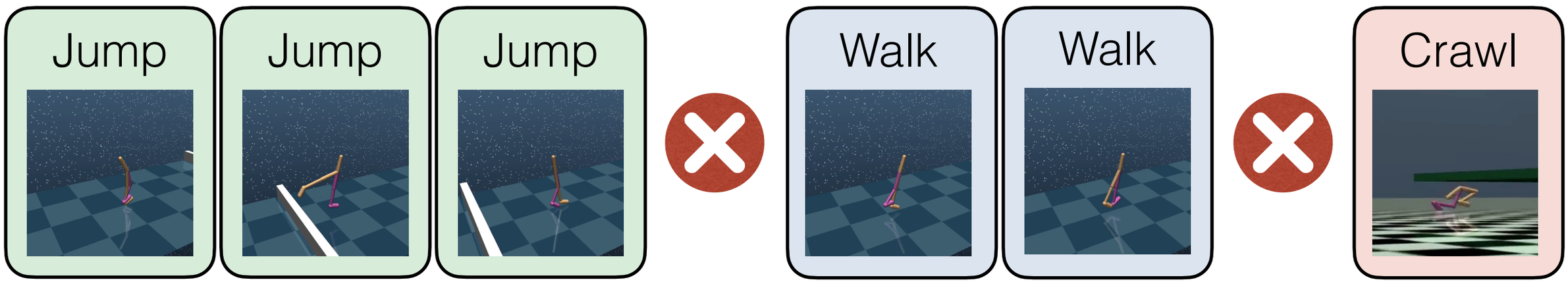


Compose Complex Skills

High-level plan



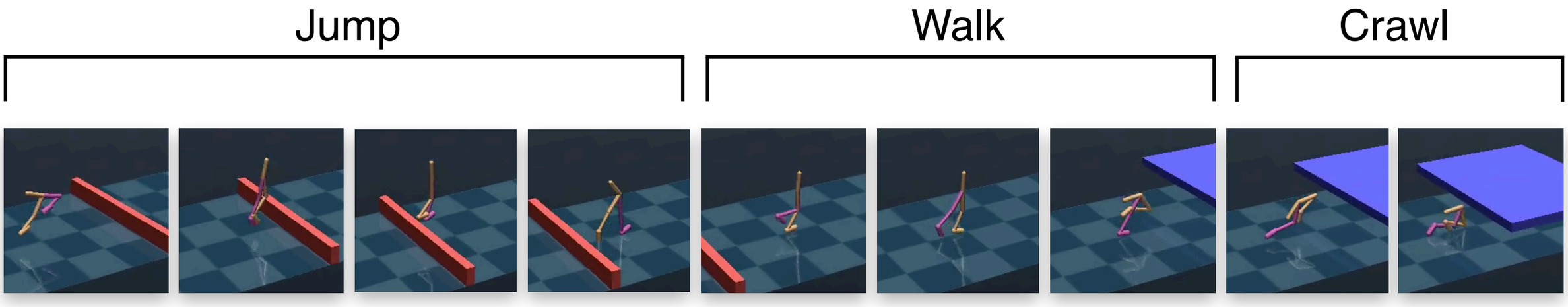
Sequentially execute corresponding policies



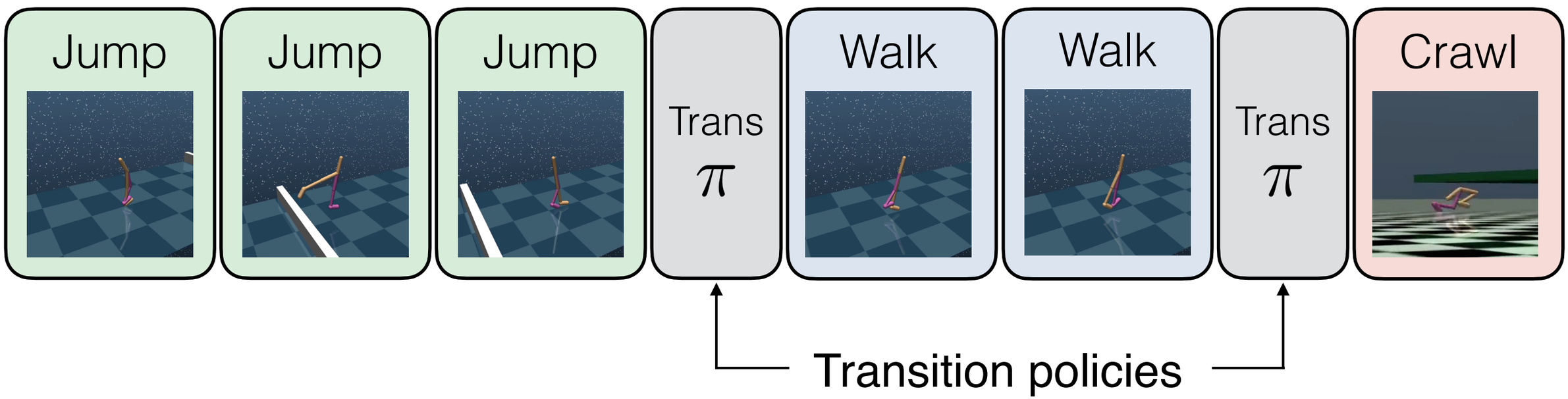
An **end state** of a previous policy might not be a good **initial state** of the following policy

Compose Complex Skills

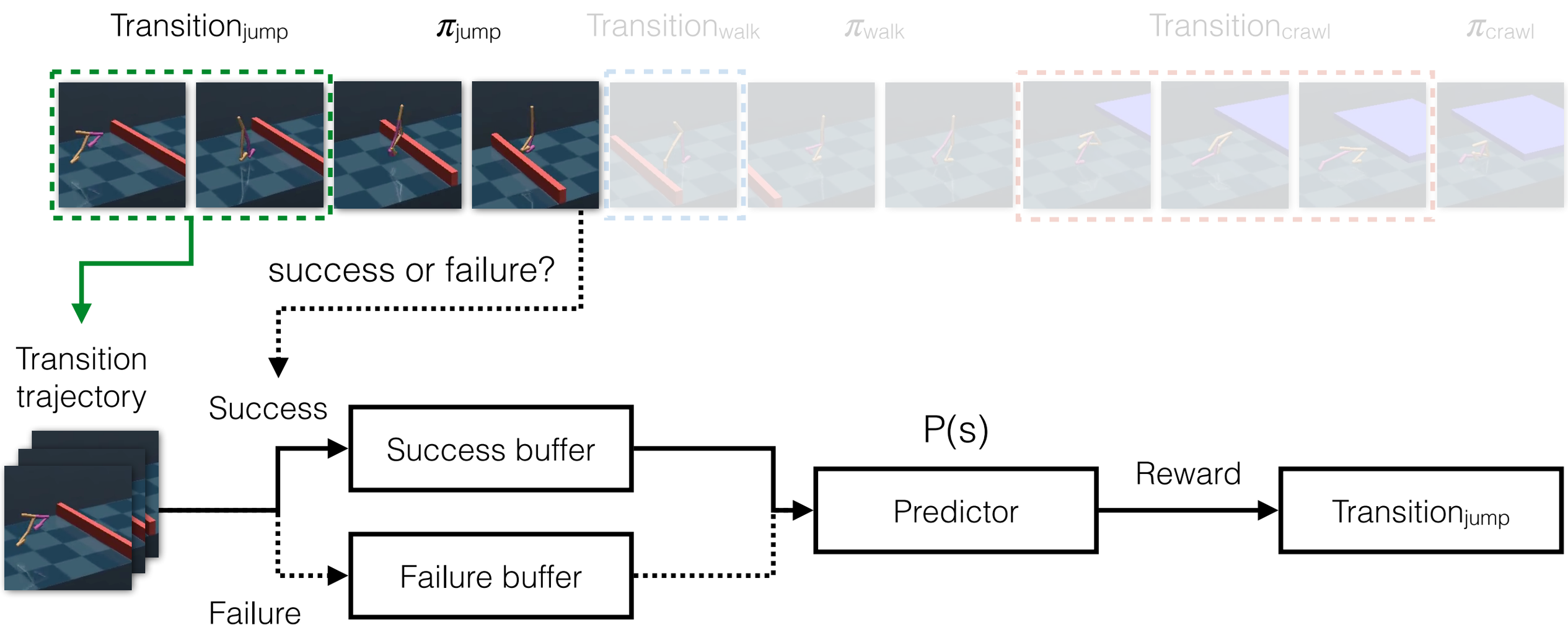
High-level plan



Sequentially execute corresponding policies

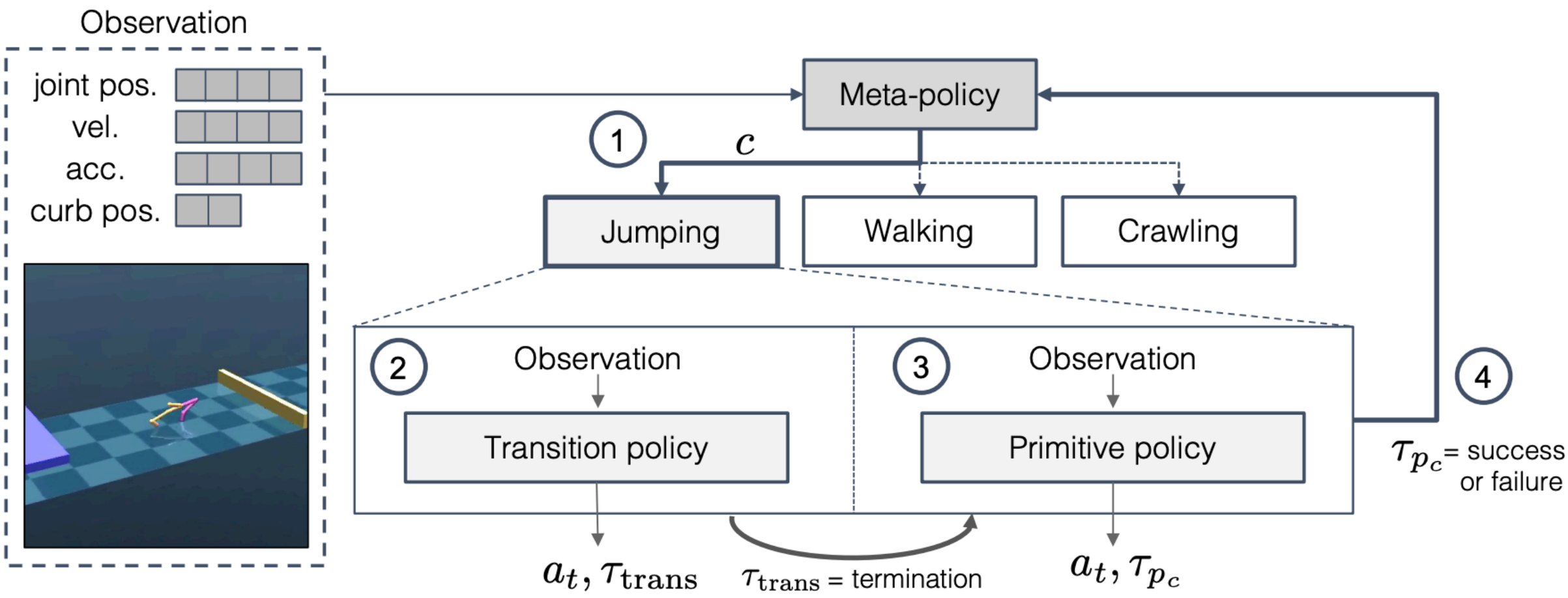


Learning Transition Policies



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

Modular Framework



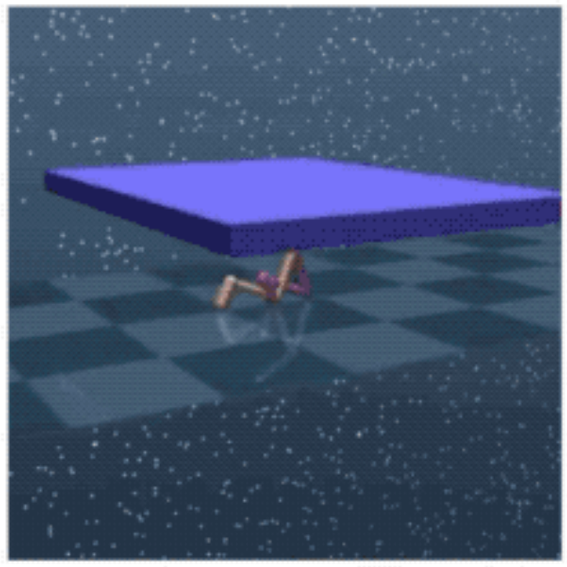
Qualitative Results

Locomotion

Crawl

Transition

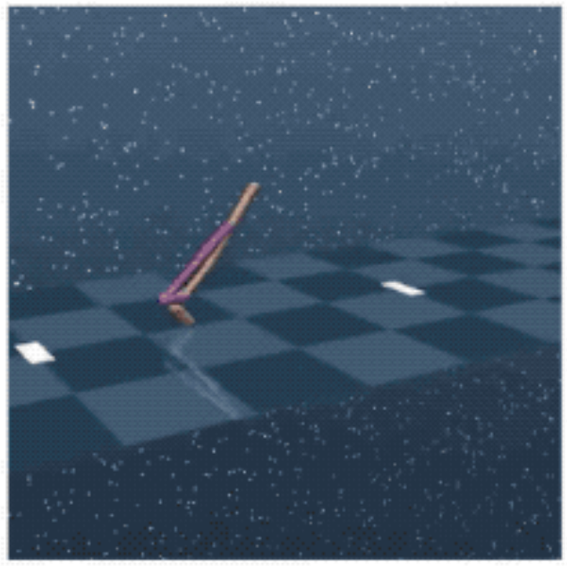
Walk



Walk Forward

Transition

Walk Backward

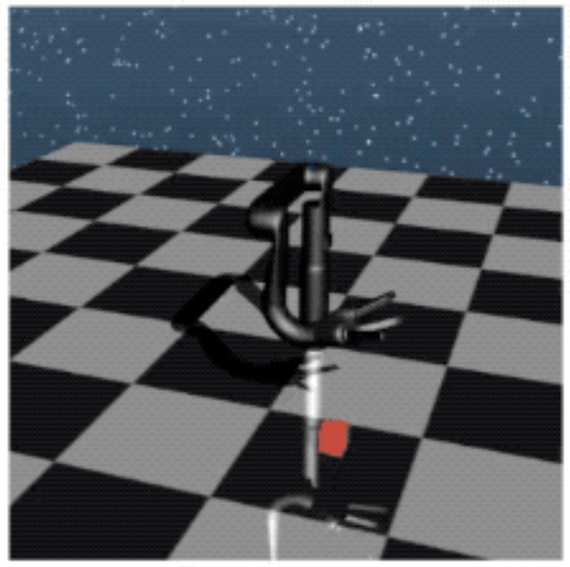


Manipulation

Pick

Transition

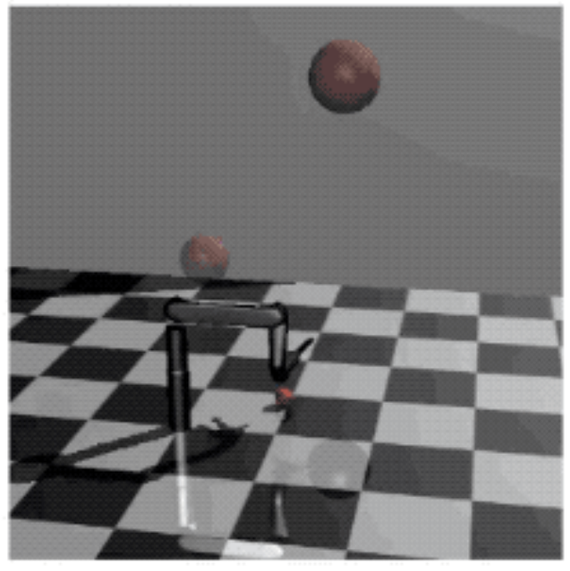
Pick



Toss

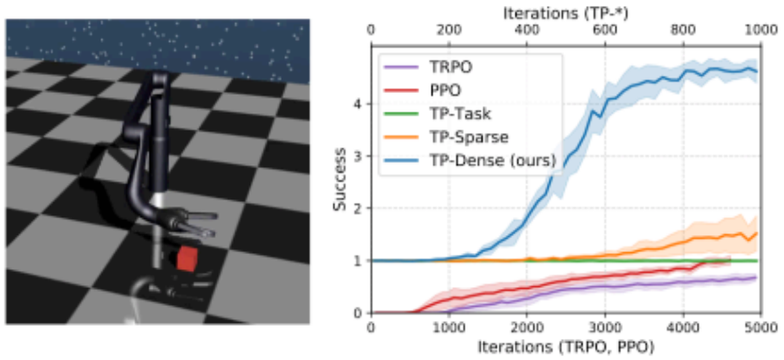
Transition

Hit

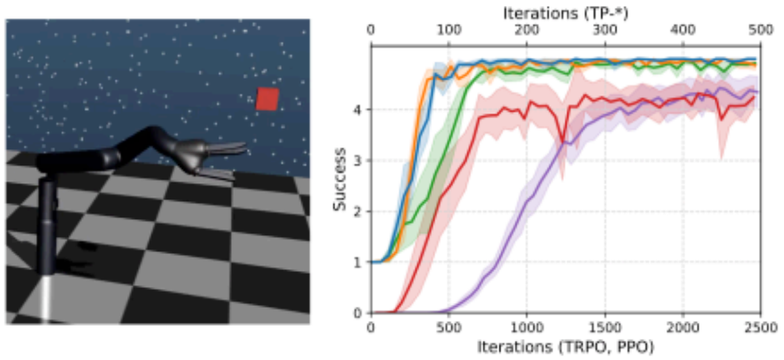


Quantitative Results

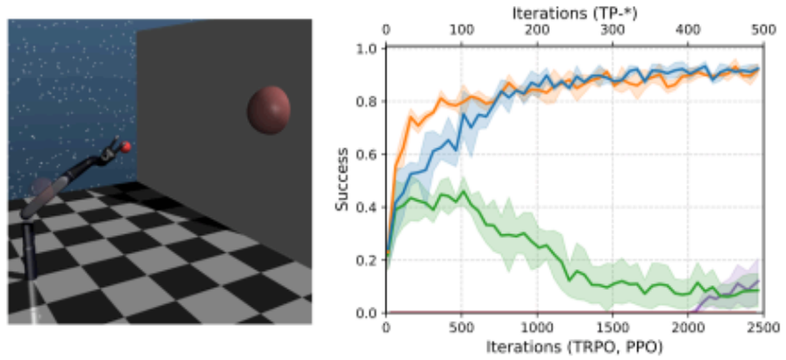
Sample Efficiency



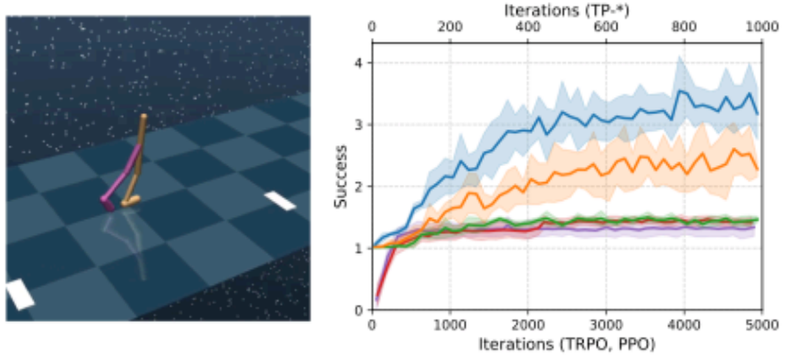
(a) Repetitive picking up



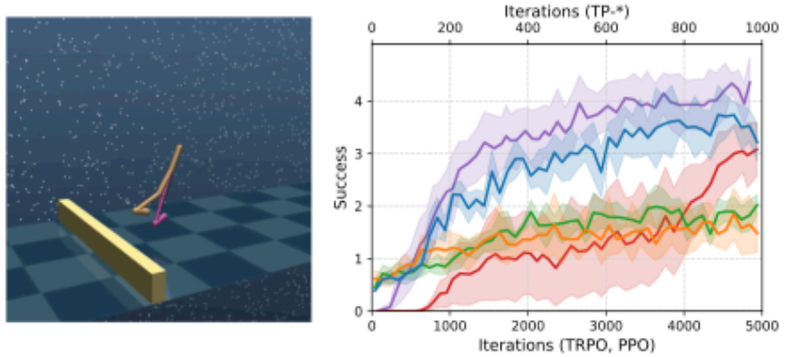
(b) Repetitive catching



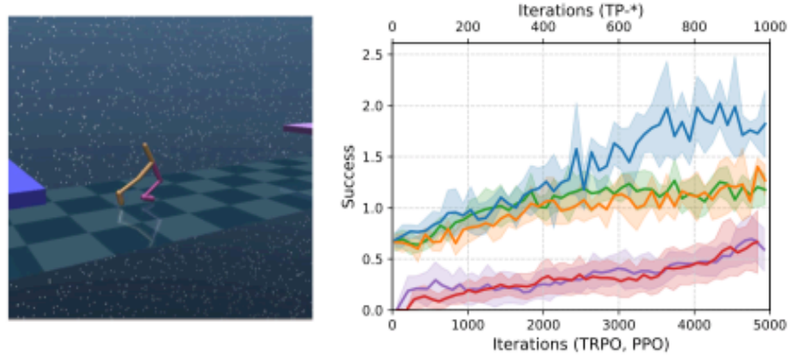
(c) Serve



(d) Patrol

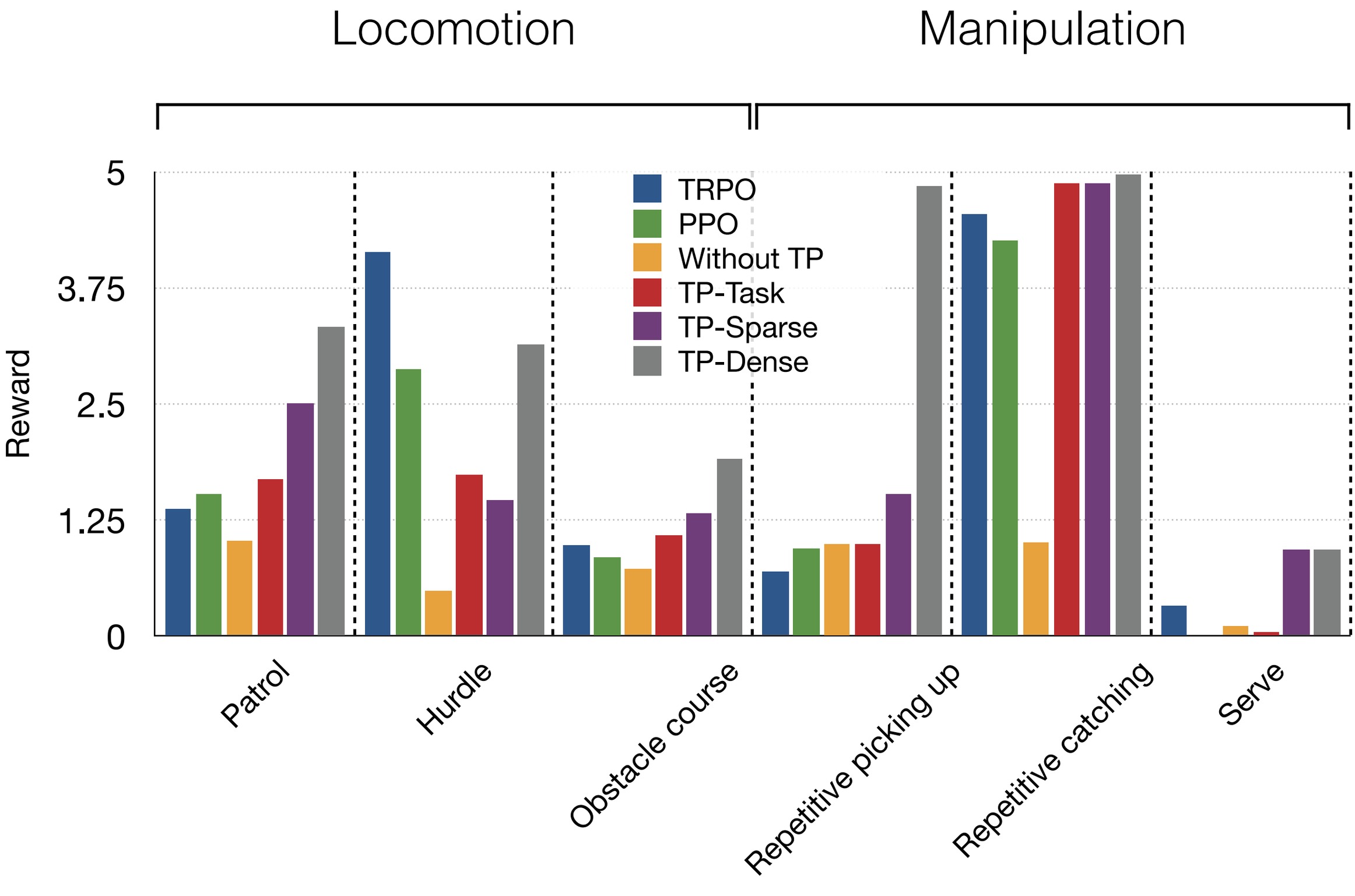


(e) Hurdle



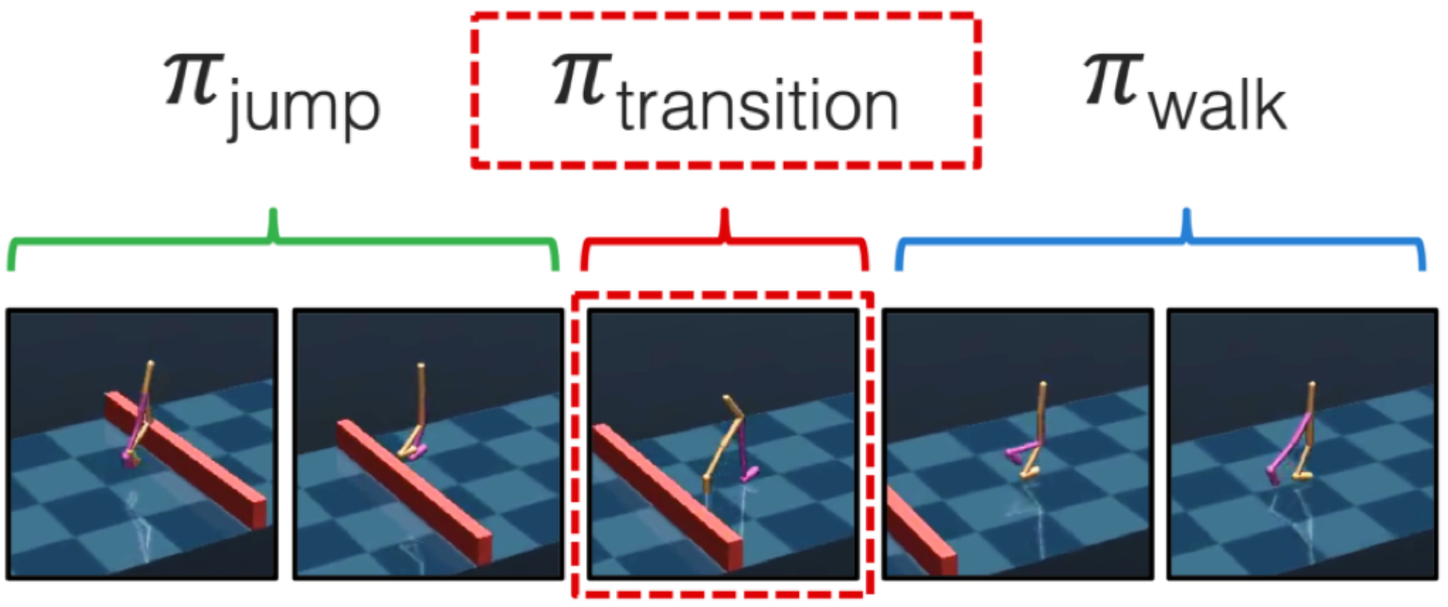
(f) Obstacle course

Quantitative Results



Takeaway

- Learning **transition policies** to smoothly compose learned skills

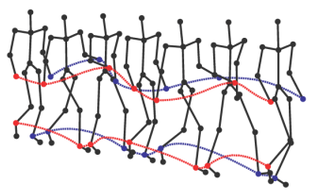


Program Inference

Skill Specification → Program → High-level Plan → Low-level Execution

Video demonstration

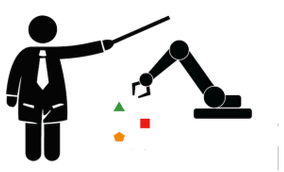
Trajectory demonstration



Language instruction

Stir-fry the onions until tender, and repeat this for garlic and carrots, if you have soy sauce, add some. Pour 2/3 cups the whisked eggs into the stir-fried and scramble.

Reward



Actions (sub-skills)

`Jump` `Crawl` `Walk`

Perceptions

`isCurb()` `isCeiling()`

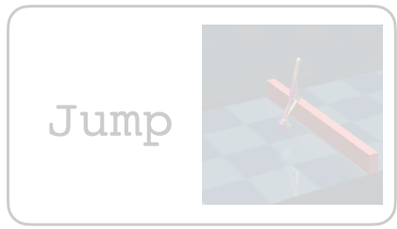
Control Flows

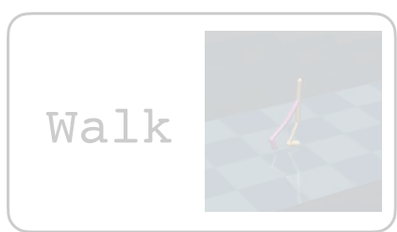
`while` `if` `else` `elseIf`

Program

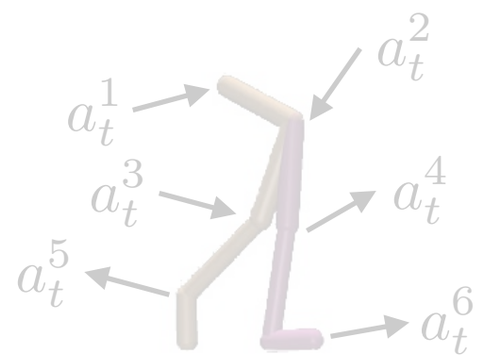
```
DEF run()
    while True:
        if isCurb():
            jump
        elseIf isCeiling():
            crawl()
        else:
            walk()
```

Observation





Low-level execution



Task Execution

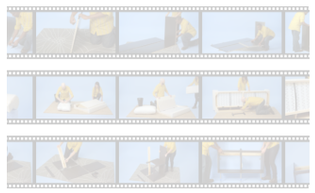
Skill Specification

Program

High-level Plan

Low-level Execution

Video demonstration



Trajectory demonstration

Language instruction

Stir-fry the onions until tender, and repeat this for garlic and carrots, if you have soy sauce, add some. Pour 2/3 cups the whisked eggs into the stir-fried and scramble.

Reward

Actions (sub-skills)

`Jump` `Crawl` `Walk`

Perceptions

`isCurb()` `isCeiling()`

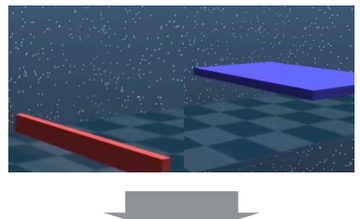
Control Flows

`while` `if` `else` `elseIf`

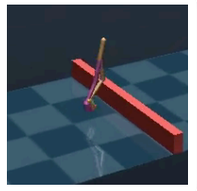
Program

```
DEF run()
    while True:
        if isCurb():
            jump
        elseIf isCeiling():
            crawl()
        else:
            walk()
```

Observation



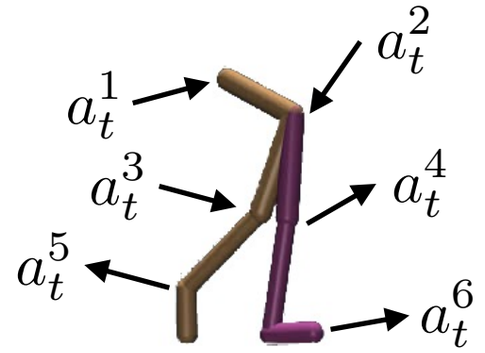
`Jump`



`Walk`

`Crawl`

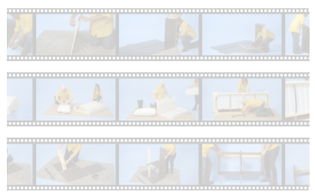
Low-level execution



Primitive Skill Acquisition

Skill Specification → Program → High-level Plan → Low-level Execution

Video demonstration



Trajectory demonstration

Language instruction

Stir-fry the onions until tender, and repeat this for garlic and carrots, if you have soy sauce, add some. Pour 2/3 cups the whisked eggs into the stir-fried and scramble.

Reward

Actions (sub-skills)

Jump

Crawl

Walk

Perceptions

isCurb()

isCeiling()

Control Flows

while | if | else | elseIf

Program

```
DEF run()
    while True:
        if isCurb():
            jump
        elseIf isCeiling():
            crawl()
        else:
            walk()
```

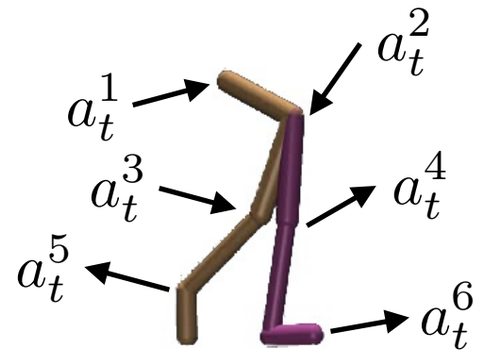
Observation

Jump

Walk

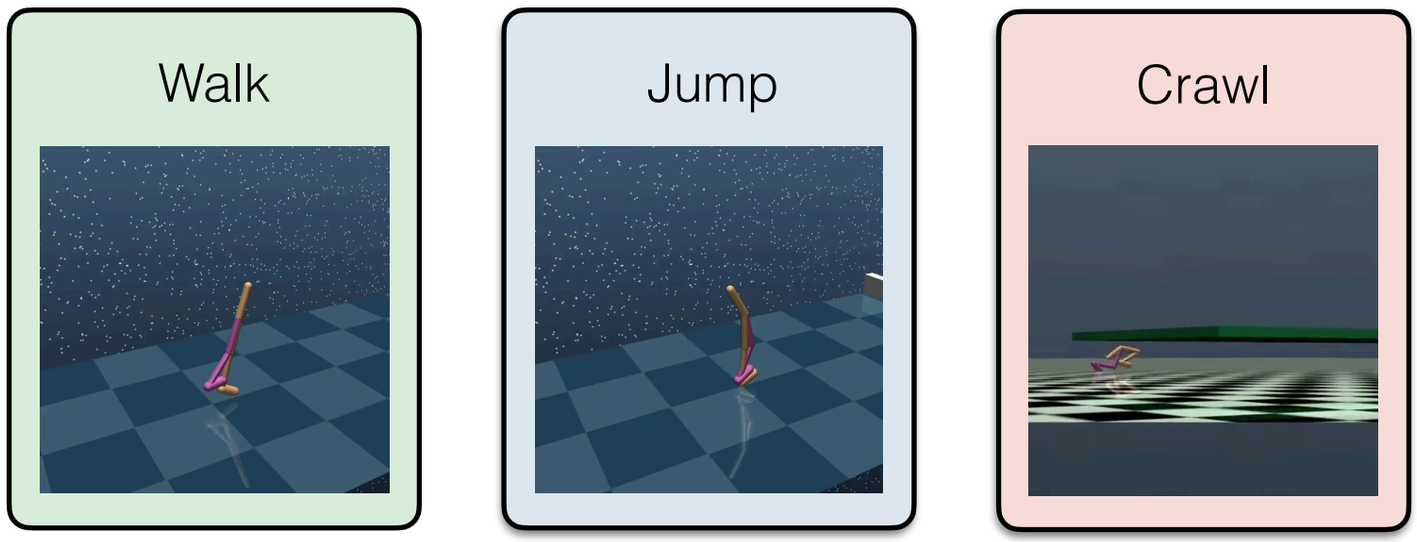
Crawl

Low-level execution



Primitive Skill Acquisition

- Goal: acquire a diverse set of primitive skills efficiently



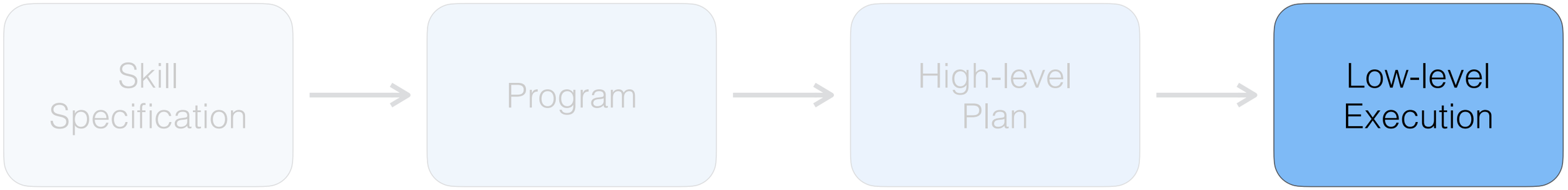
- Key directions

Meta-learning
Meta-RL

Learning from
experts

Toward Multimodal Model-Agnostic Meta-Learning

Meta-learning Workshop at NeurIPS 2018

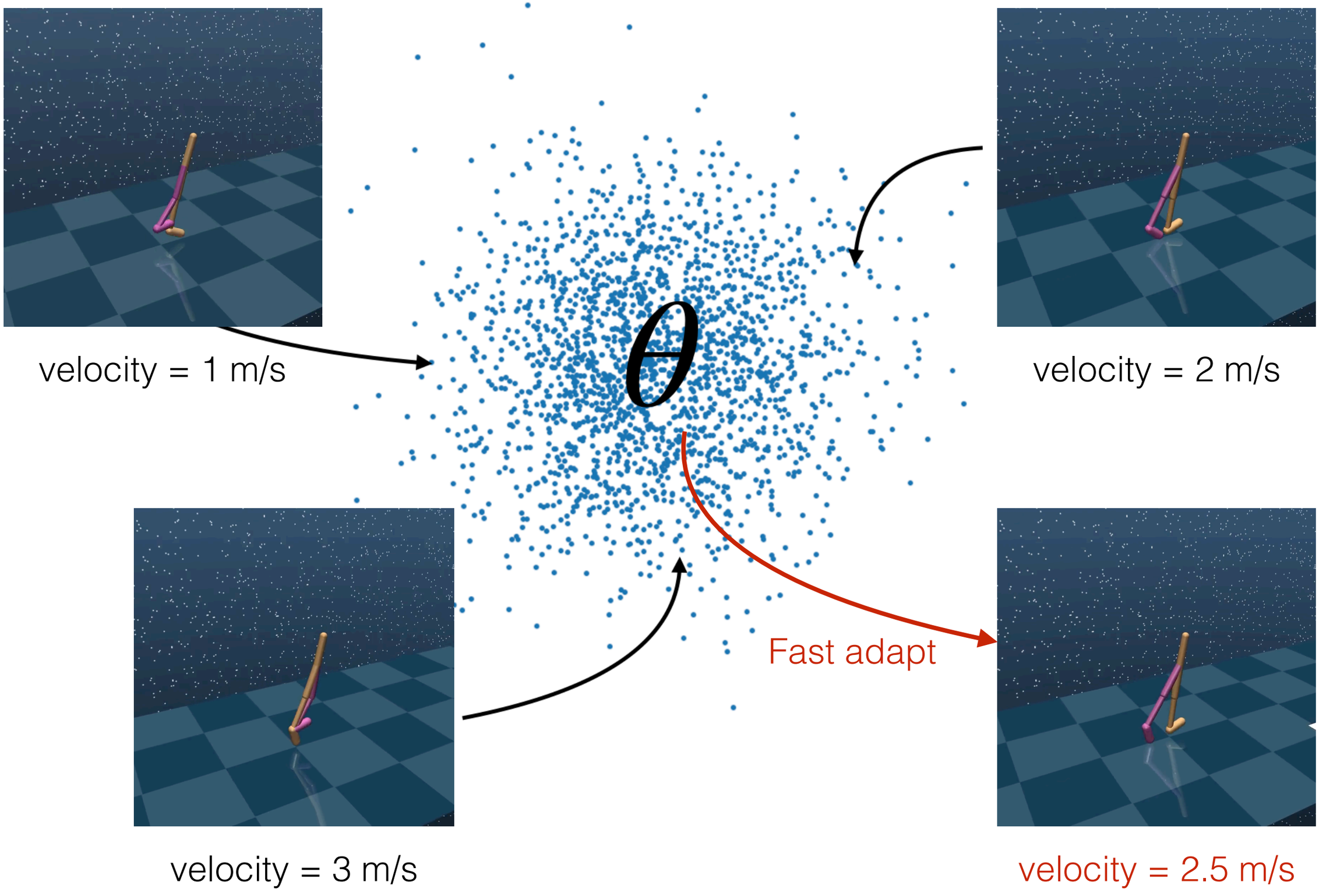



Risto Vuorio


Hexiang Hu

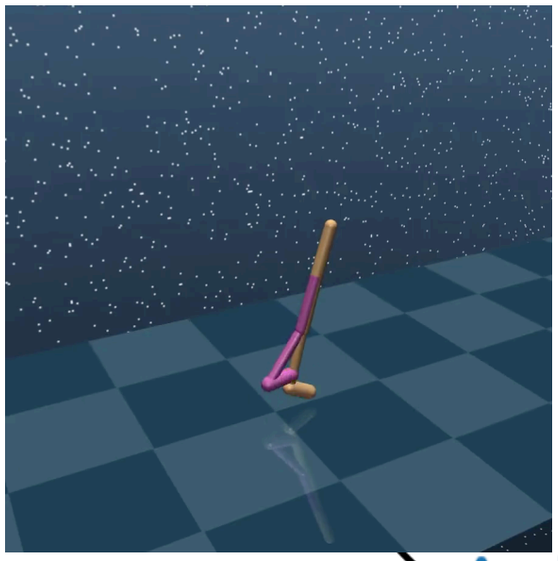

Joseph J. Lim

Model-Agnostic Meta-Learning (MAML)

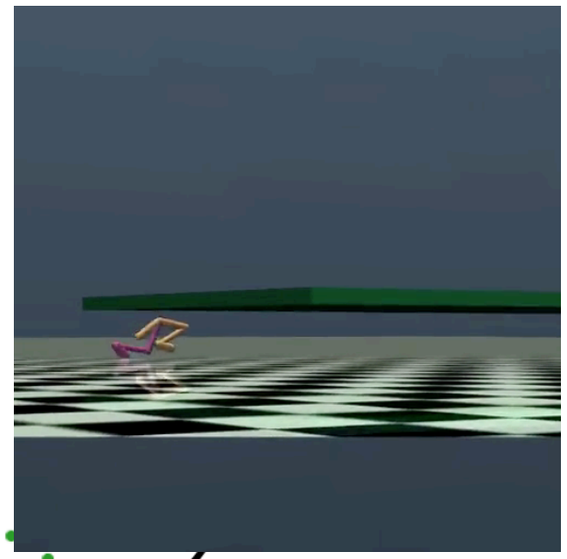


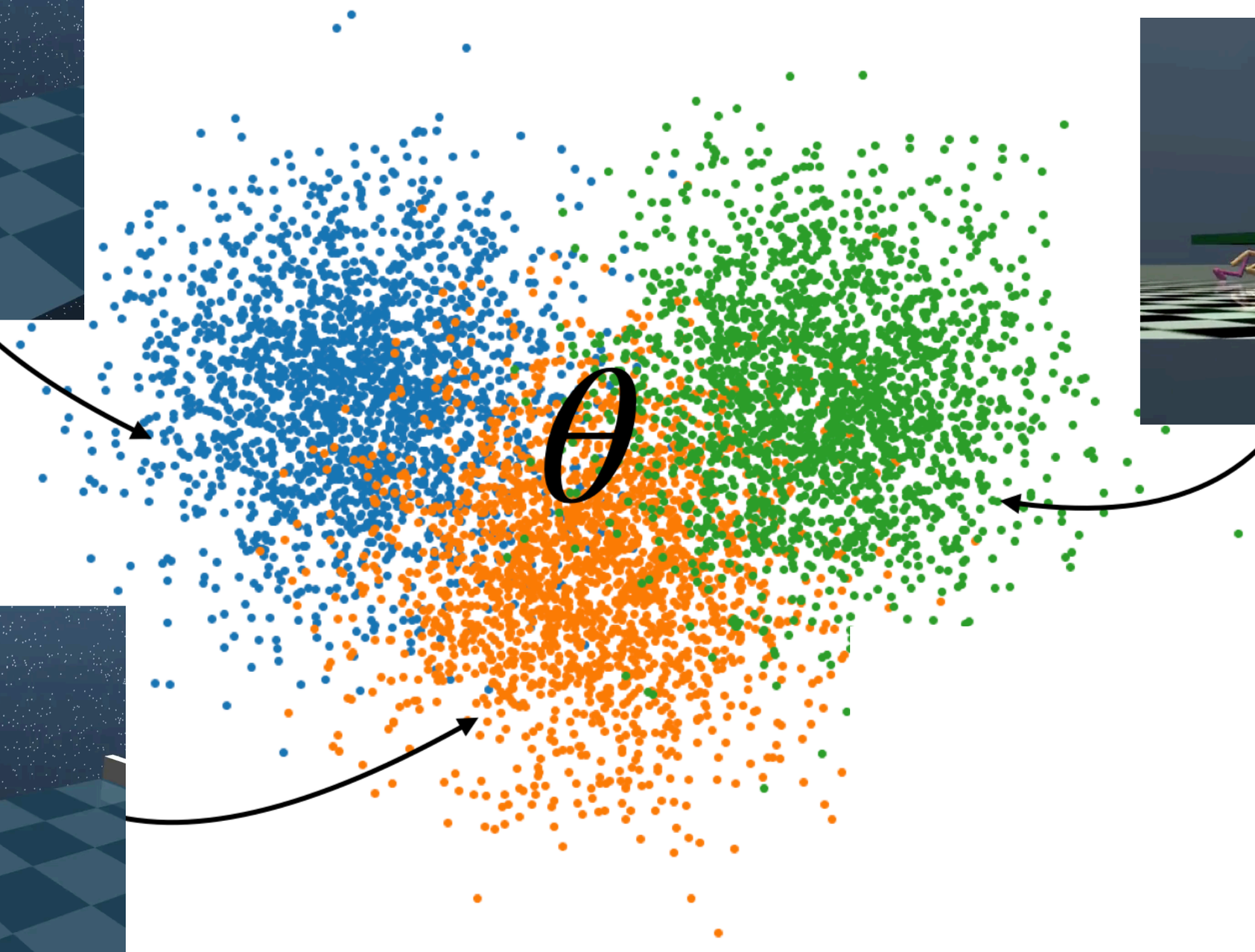
Multimodal Task Distribution

Walk

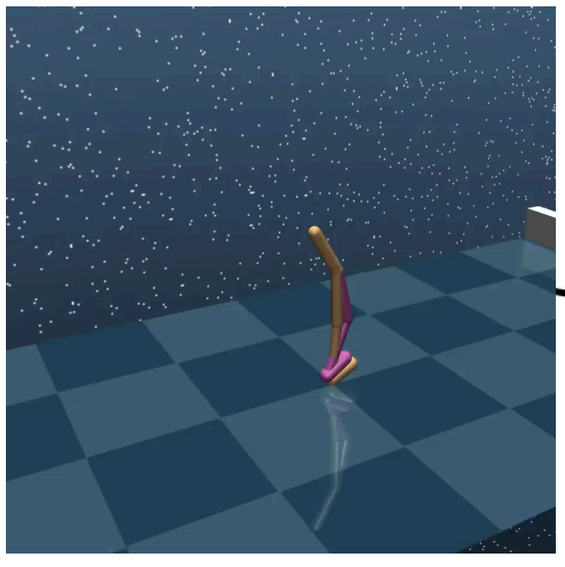


Crawl



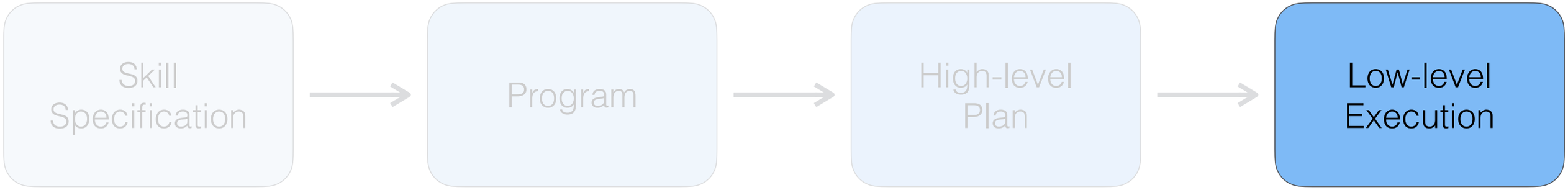


Jump



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

NeurIPS 2019 (Spotlight)

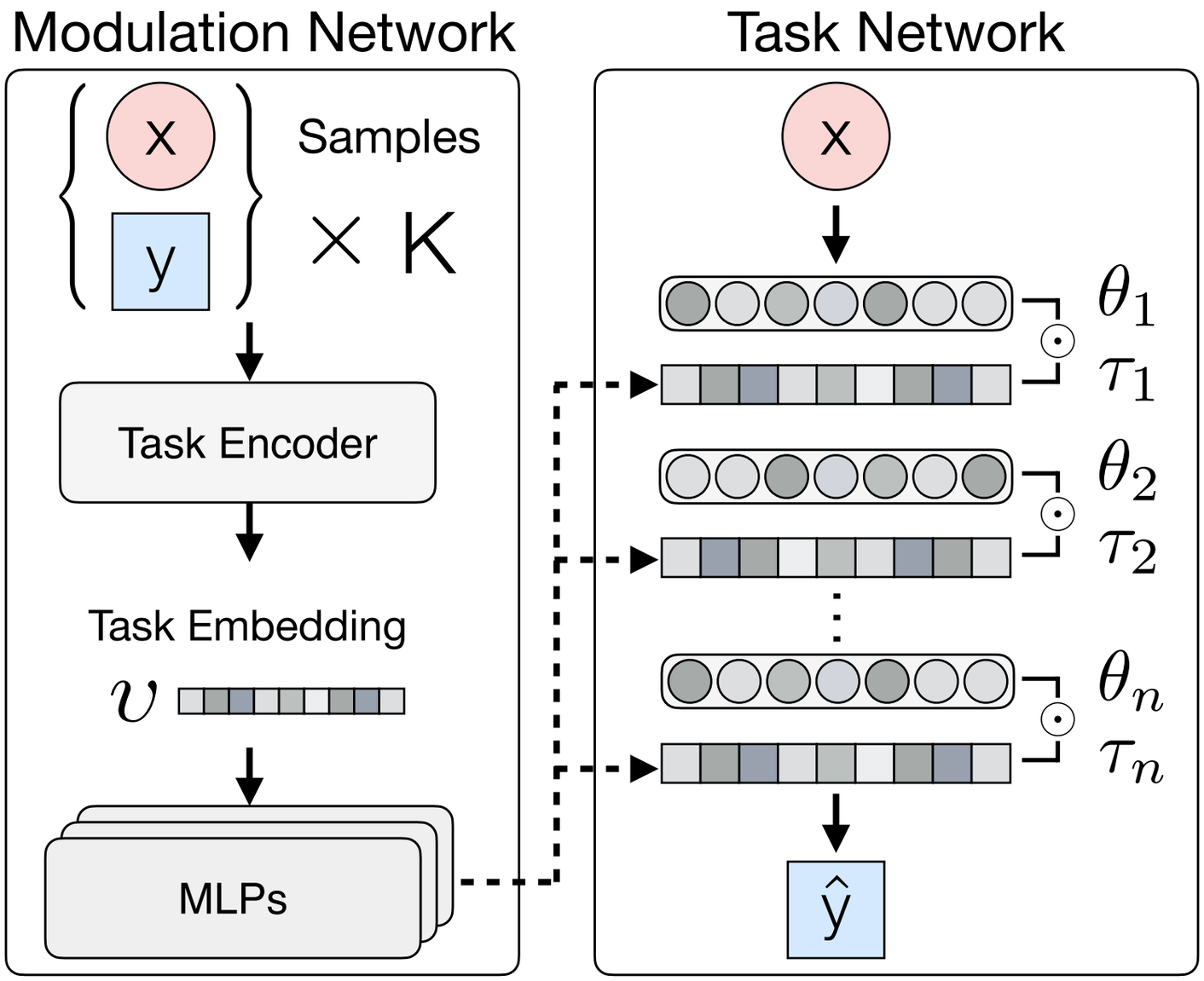



Risto Vuorio


Hexiang Hu


Joseph J. Lim

Multimodal Model-Agnostic Meta-Learning (MMAML)



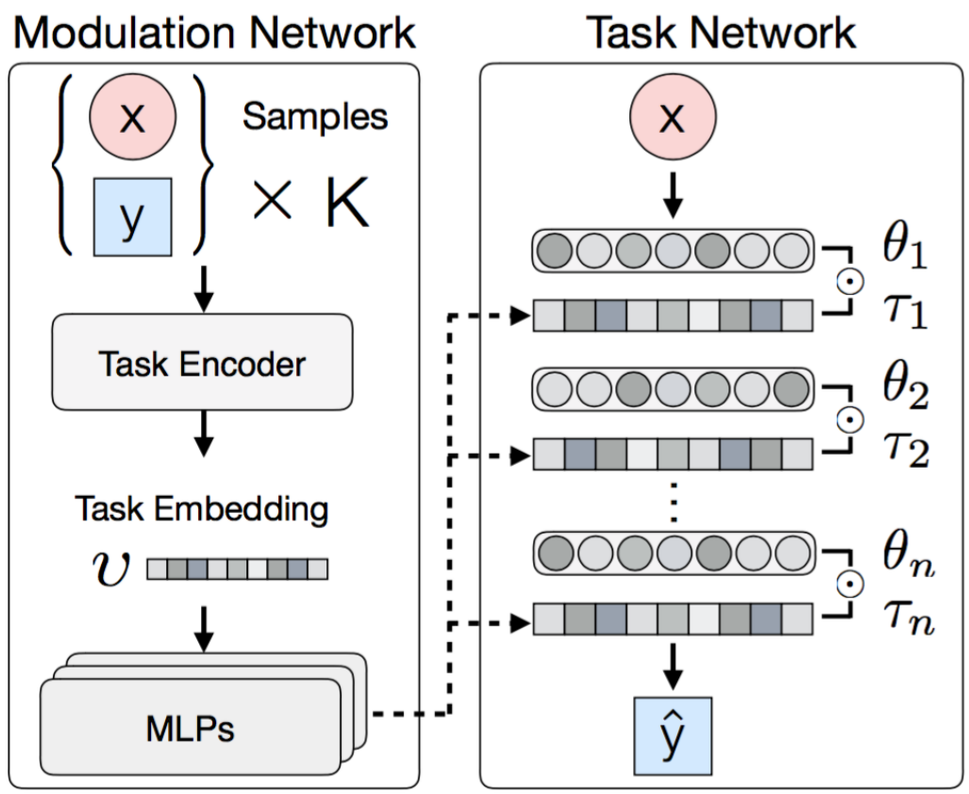
Training Algorithm

Outer loop

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

Inner loop

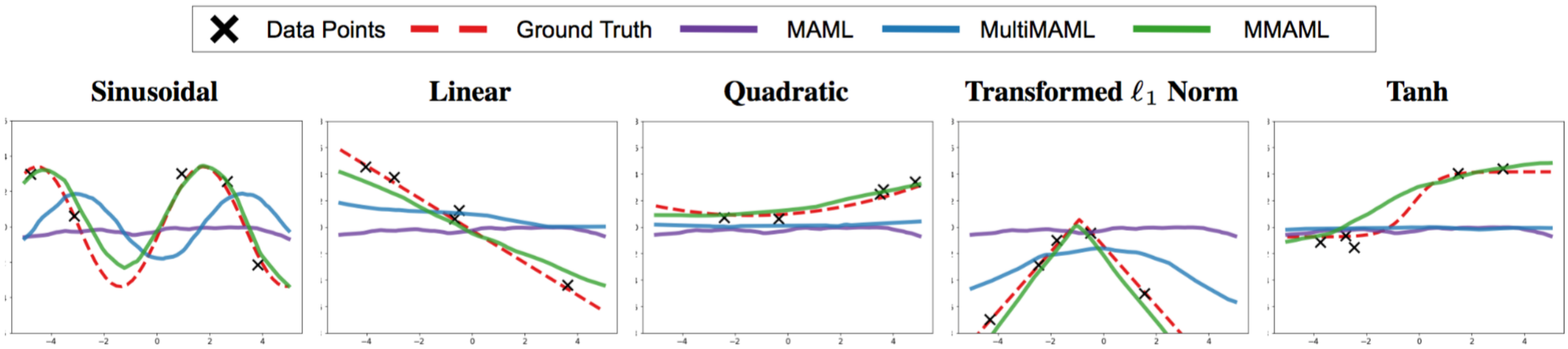
- Task network: fast adapt through gradient updates θ



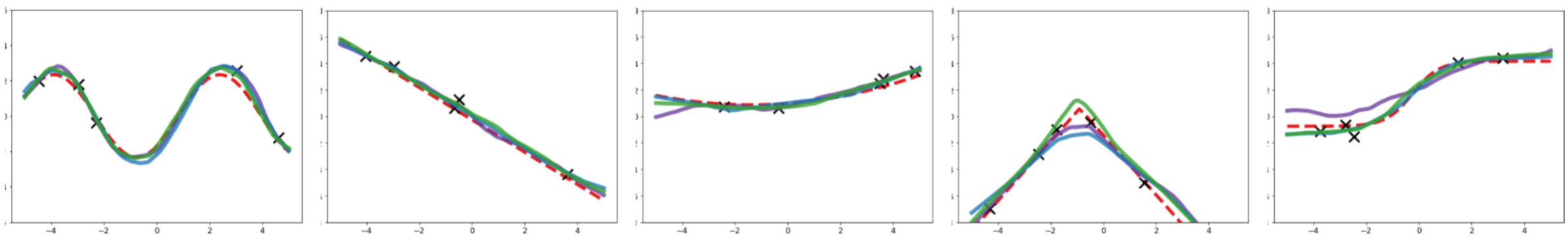
Algorithm 1 MMAML META-TRAINING PROCEDURE.

Input: Task distribution $P(\mathcal{T})$, Hyper-parameters α and β
Randomly initialize θ and ω.
while not DONE **do**
  Sample batches of tasks $\mathcal{T}_j \sim P(\mathcal{T})$
  for all j **do**
    Infer $\upsilon = h(\{x, y\}_K; \omega_h)$ with K samples from $\mathcal{D}^{\text{train}}_{\mathcal{T}_j}$.
    Generate parameters $\tau = \{g_i(\upsilon; \omega_g) \mid i = 1, \cdots, N\}$ to modulate each block of the task network f.
    Evaluate $\nabla_\theta \mathcal{L}_{\mathcal{T}_j}(f(x; \theta, \tau); \mathcal{D}^{\text{train}}_{\mathcal{T}_j})$ w.r.t the K samples
    Compute adapted parameter with gradient descent: $\theta'_{\mathcal{T}_j} = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_j}(f(x; \theta, \tau); \mathcal{D}^{\text{train}}_{\mathcal{T}_j})$
  end for
  Update θ with $\beta \nabla_\theta \sum_{T_j \sim P(\mathcal{T})} \mathcal{L}_{\mathcal{T}_j}(f(x; \theta', \tau); \mathcal{D}^{\text{val}}_{\mathcal{T}_j})$
  Update ω_g with $\beta \nabla_{\omega_g} \sum_{T_j \sim P(\mathcal{T})} \mathcal{L}_{\mathcal{T}_j}(f(x; \theta', \tau); \mathcal{D}^{\text{val}}_{\mathcal{T}_j})$
  Update ω_h with $\beta \nabla_{\omega_h} \sum_{T_j \sim P(\mathcal{T})} \mathcal{L}_{\mathcal{T}_j}(f(x; \theta', \tau); \mathcal{D}^{\text{val}}_{\mathcal{T}_j})$
end while

Regression



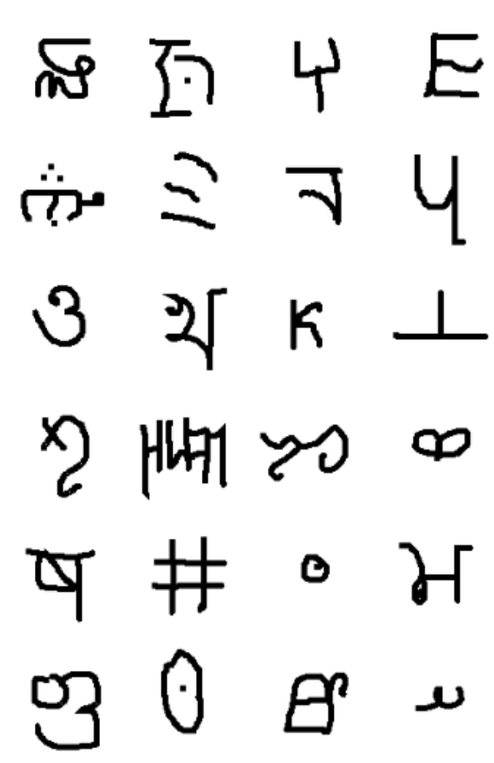
(a) MMAML **post modulation** vs. other prior models

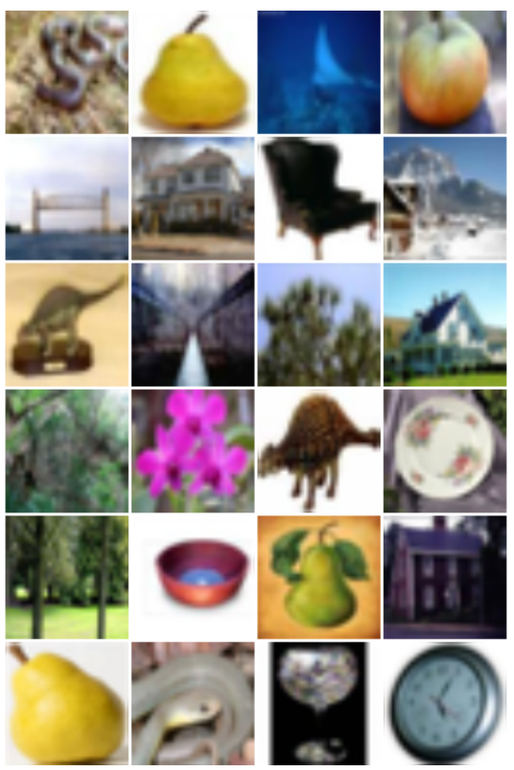


(b) MMAML **post adaptation** vs. other posterior models

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**

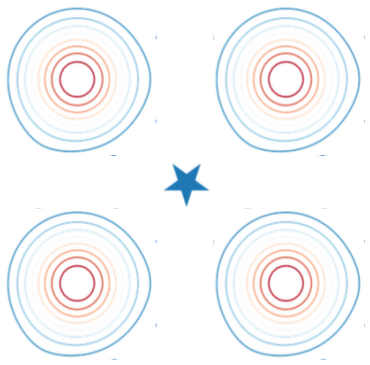
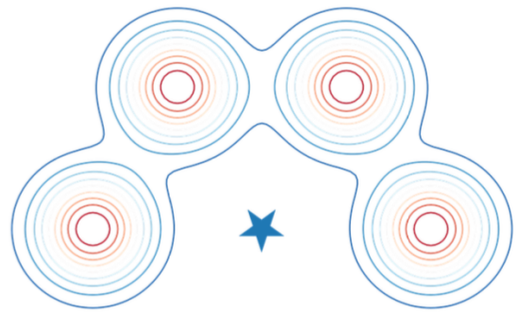
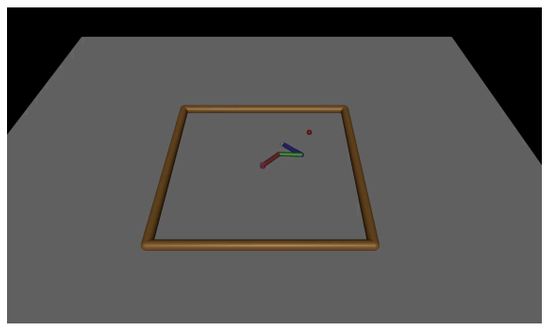
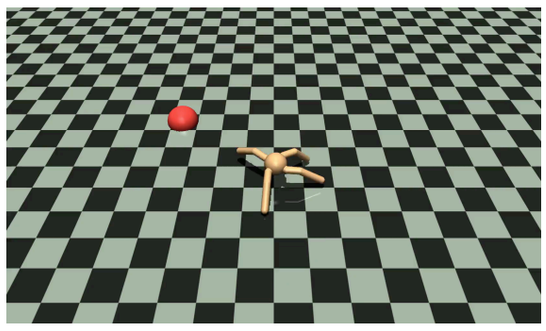
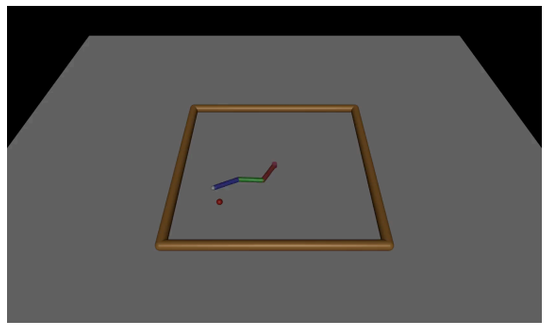
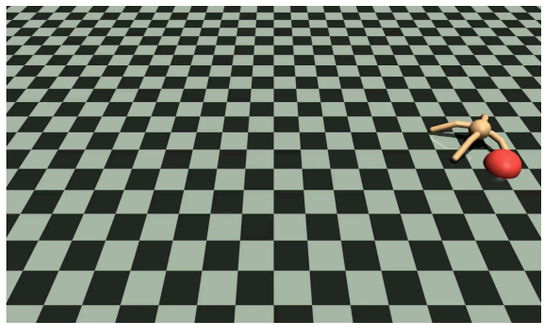
Image Classification



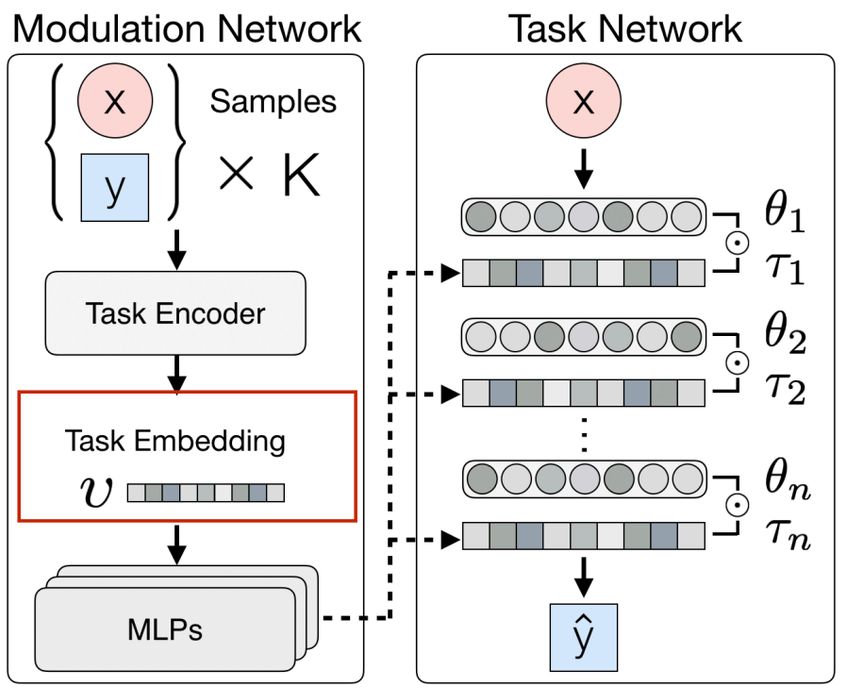


Method & Setup	**2 Modes**			**3 Modes**			**5 Modes**			
Way	5-way		20-way	5-way		20-way	5-way		20-way	
Shot	1-shot	5-shot	1-shot	1-shot	5-shot	1-shot	1-shot	5-shot	1-shot	3-shot
MAML [1]	66.80%	77.79%	44.69%	54.55%	67.97%	28.22%	66.00%	70.87%	44.69%	56.57%
Multi-MAML	66.85%	73.07%	**53.15%**	55.90%	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%**

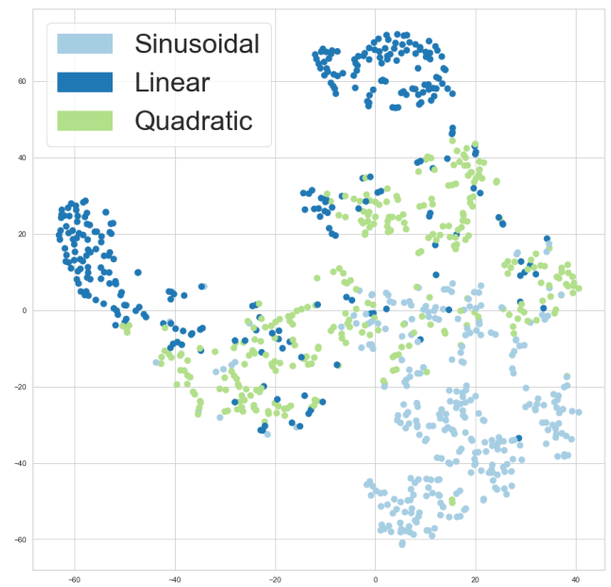
Reinforcement Learning

	Reacher	Ant
Goal modes		
ProMP		
Ours		

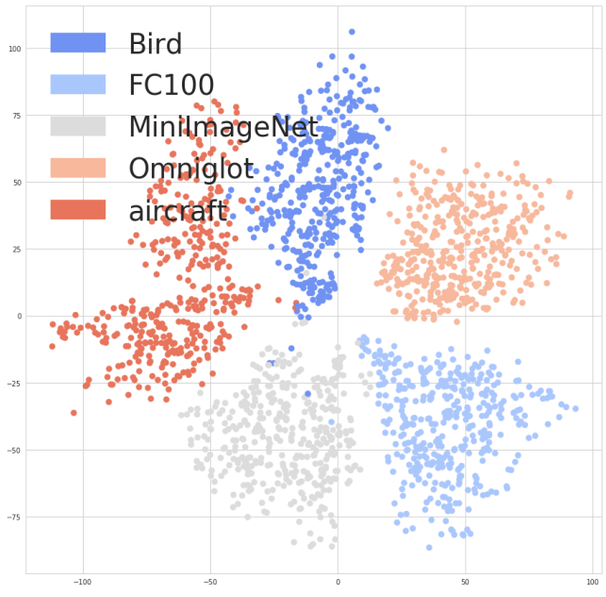
Learned Task Embedding (tSNE plot)



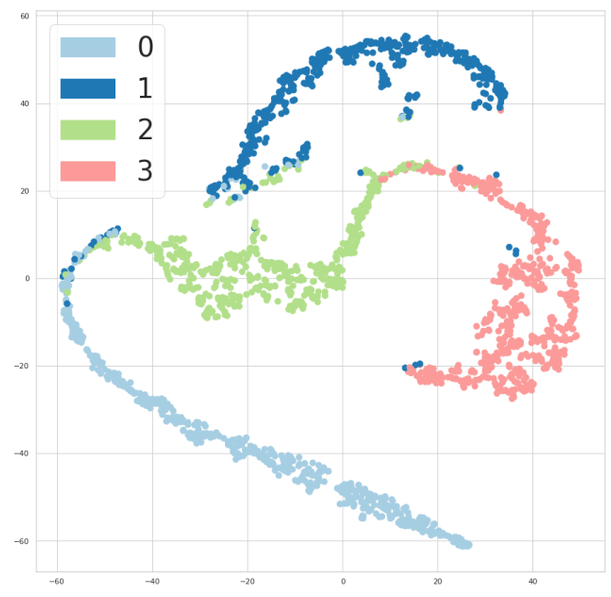
3-mode Regression



5-mode Classification

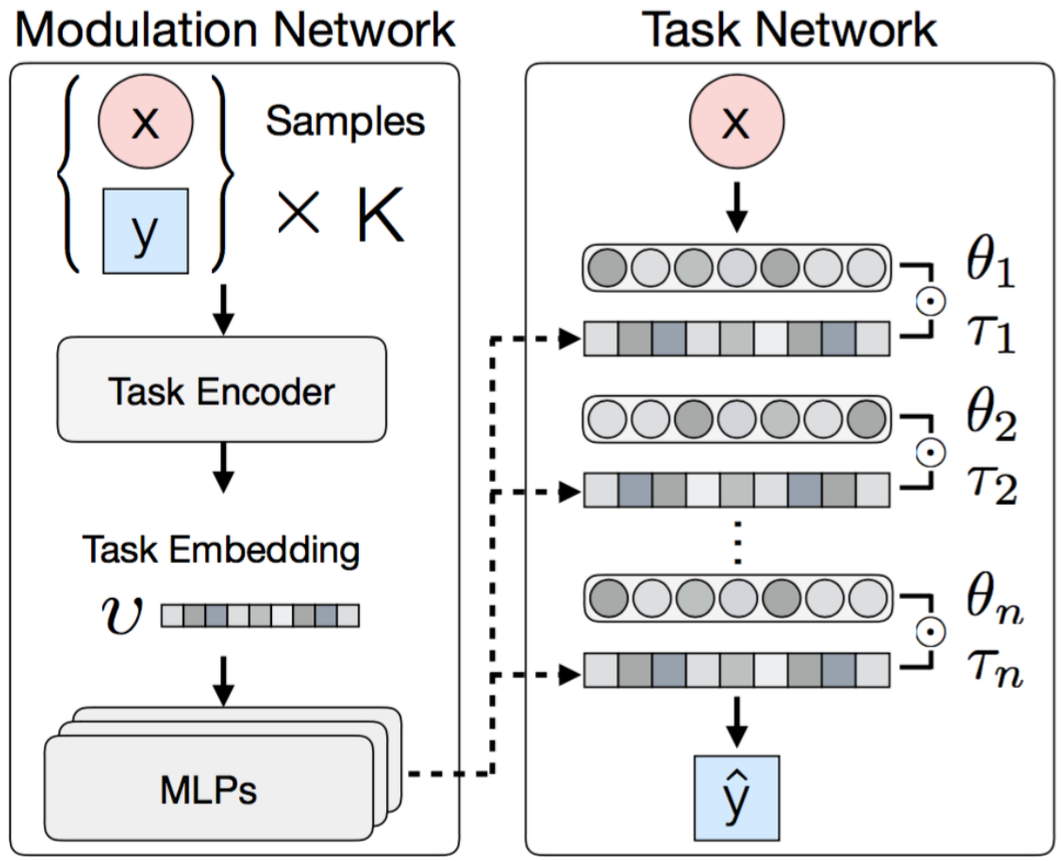


4-mode Reacher



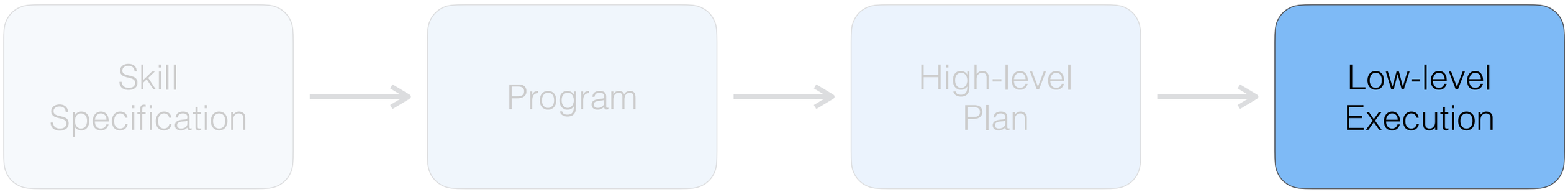
Takeaway

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



Skill-based Meta-Reinforcement Learning

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



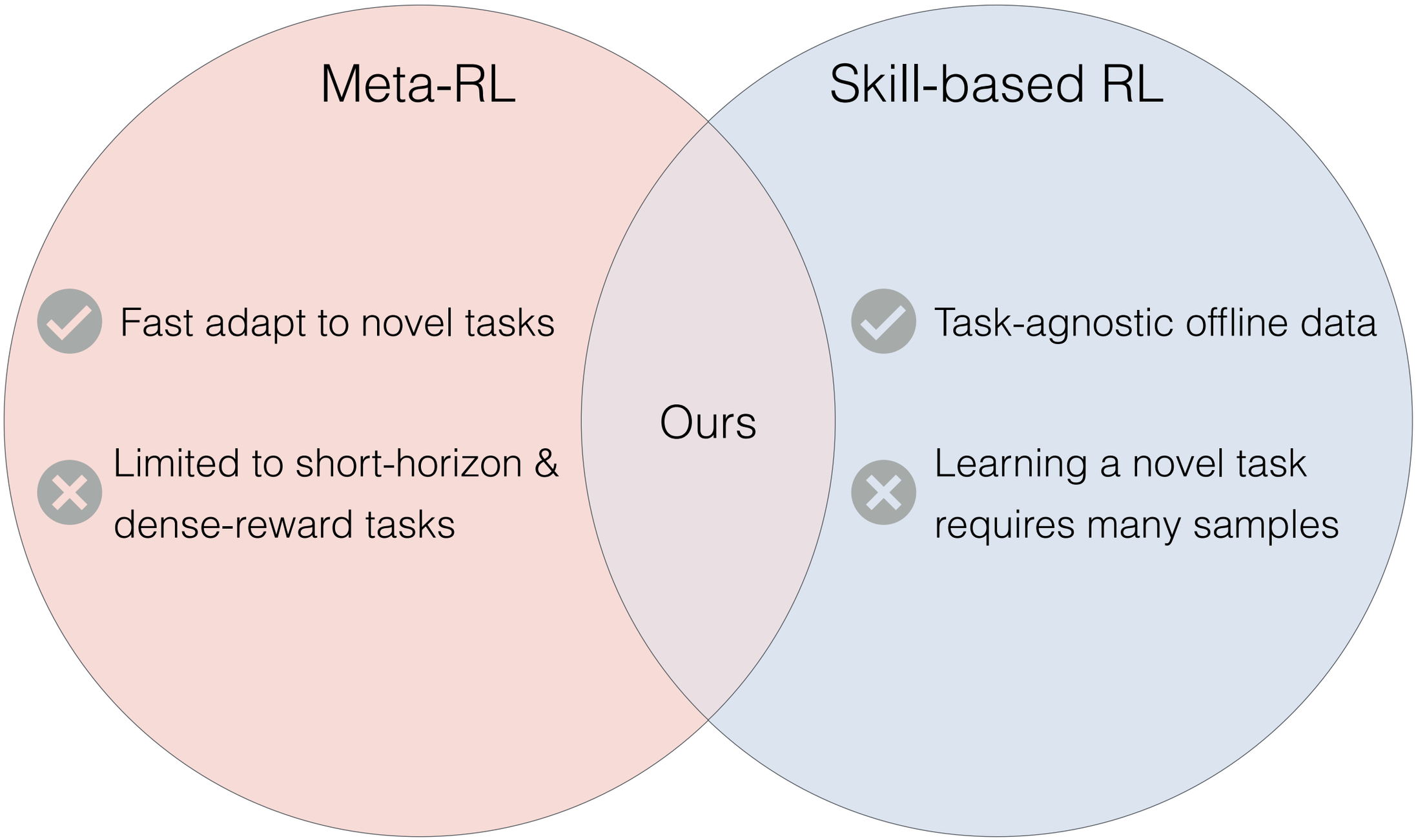

Taewook Nam


Karl Pertsch

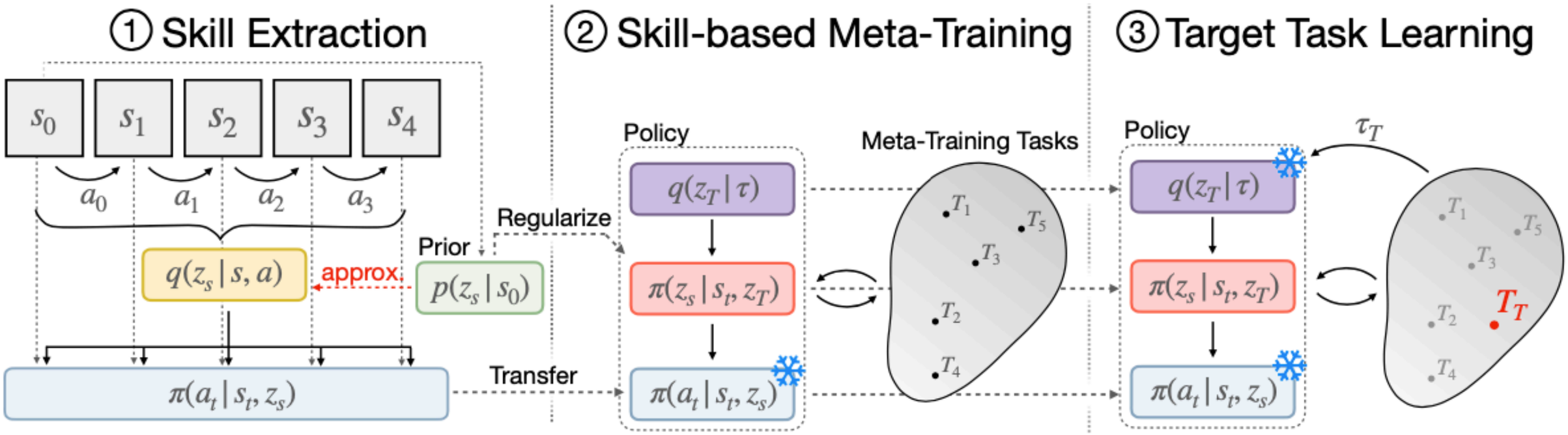

Sung Ju Hwang


Joseph J. Lim

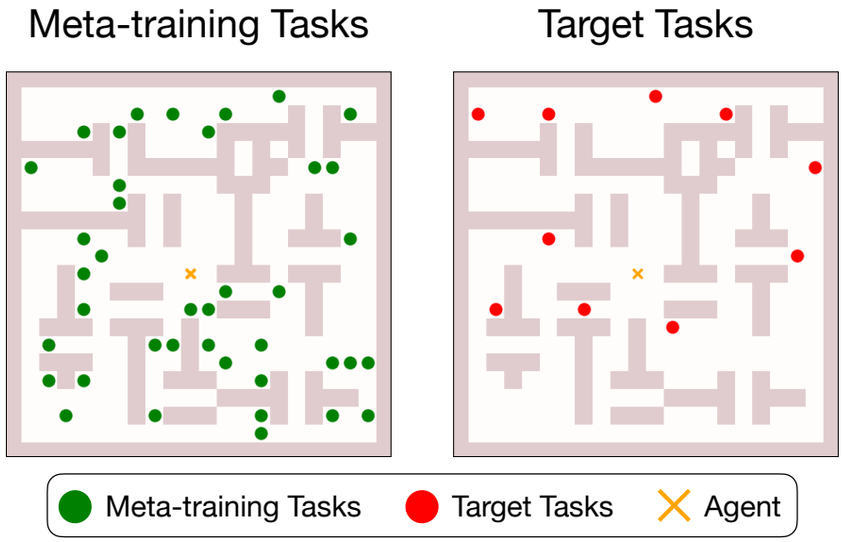
Meta-RL with Skills



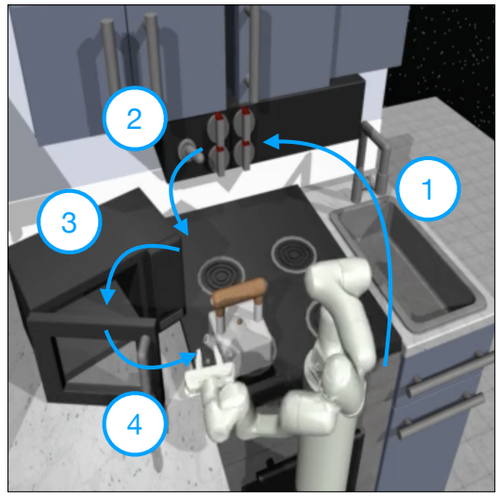
SiMPL: **S**kill-based **M**eta **P**olicy **L**earning

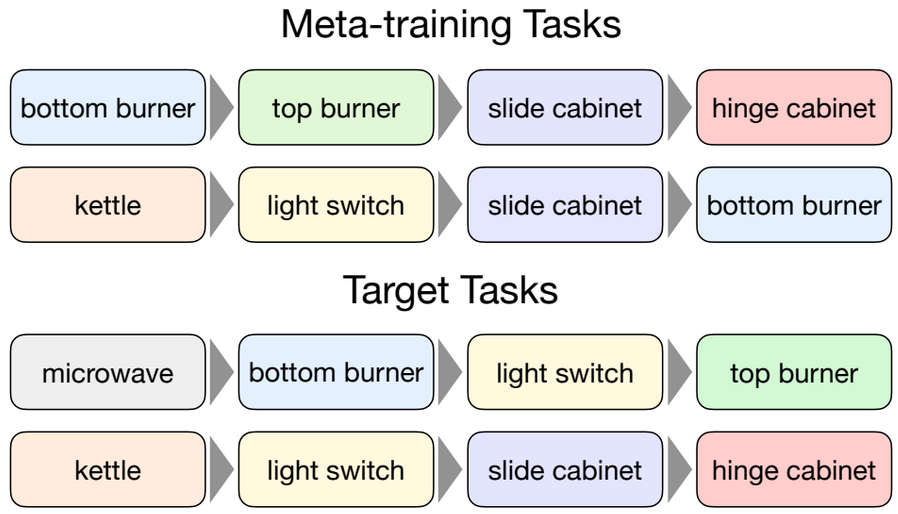


Environments



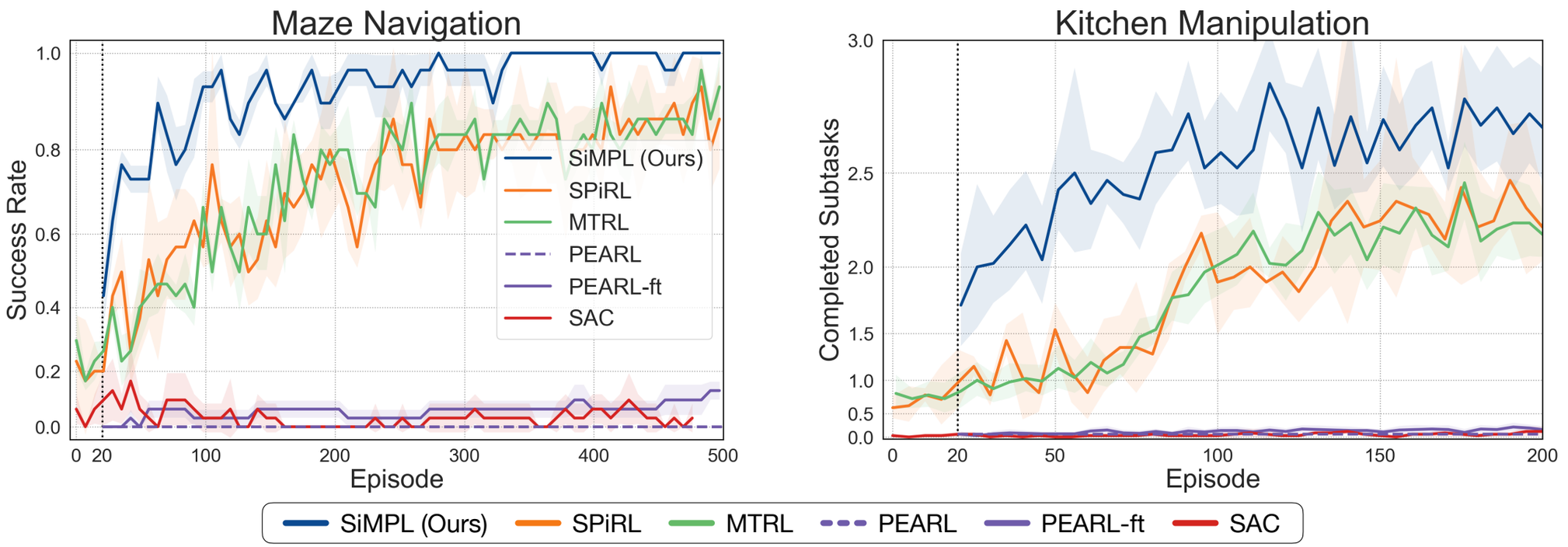
(a) Maze Navigation

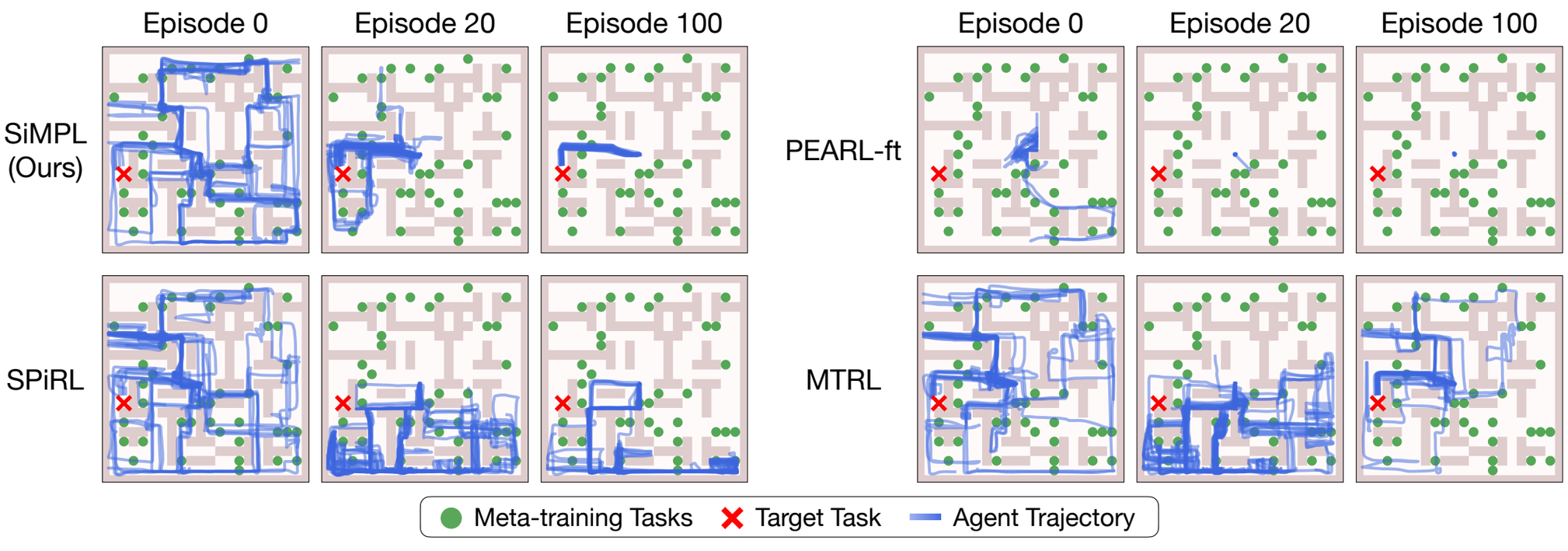




(b) Kitchen Manipulation

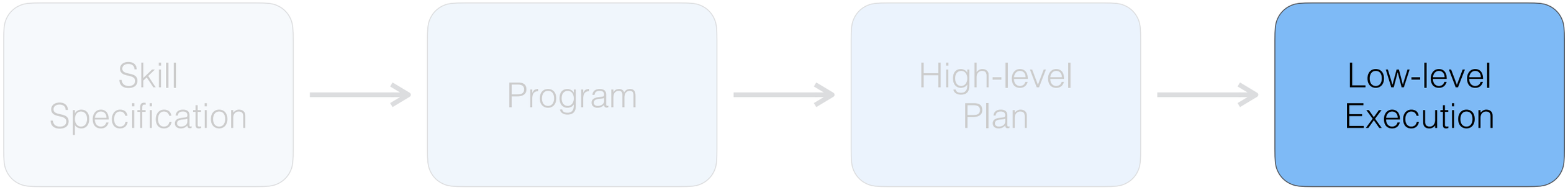
Results


Maze Navigation
Success Rate
1.0
0.8
0.6
0.4
0.2
0.0
0
20
100
200
300
400
500
Episode
SiMPL (Ours)
SPiRL
MTRL
PEARL
PEARL-ft
SAC
Kitchen Manipulation
Completed Subtasks
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
20
50
100
150
200
Episode
SiMPL (Ours)
SPiRL
MTRL
PEARL
PEARL-ft
SAC


Episode 0
Episode 20
Episode 100
SiMPL (Ours)
SPiRL
PEARL-ft
MTRL
Meta-training Tasks
Target Task
Agent Trajectory

Generalizable Imitation Learning from Observation via Inferring Goal Proximity

NeurIPS 2021

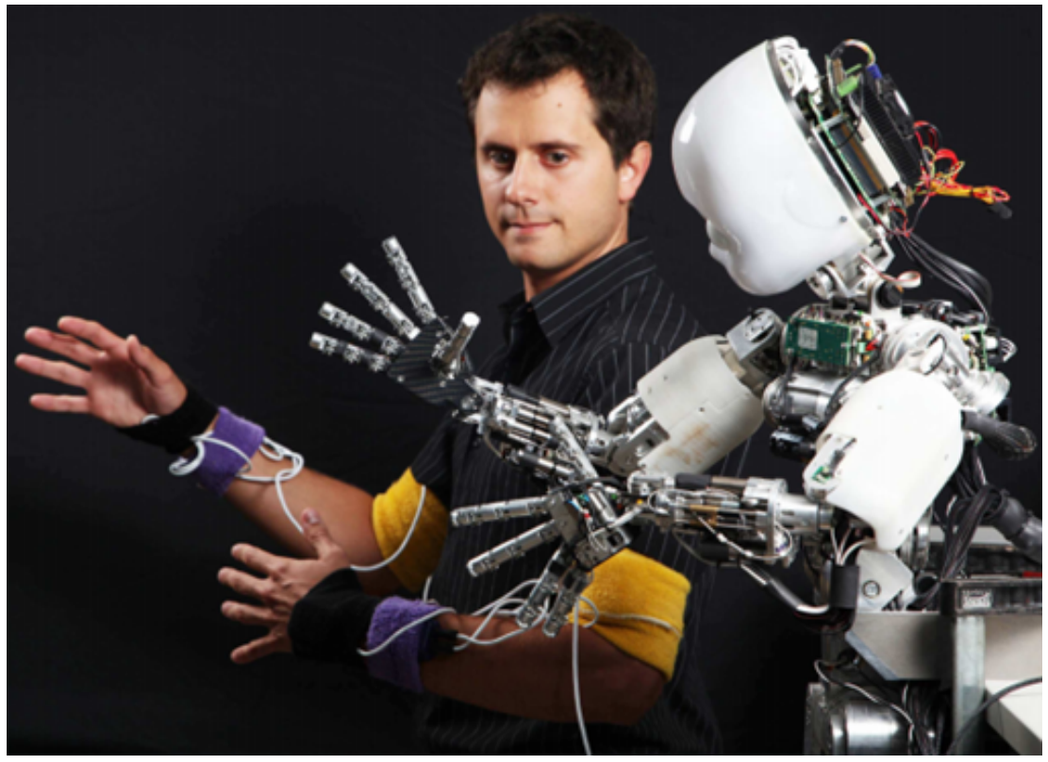



Youngwoon Lee


Andrew Szot


Joseph J. Lim

Learning from **Demonstration**

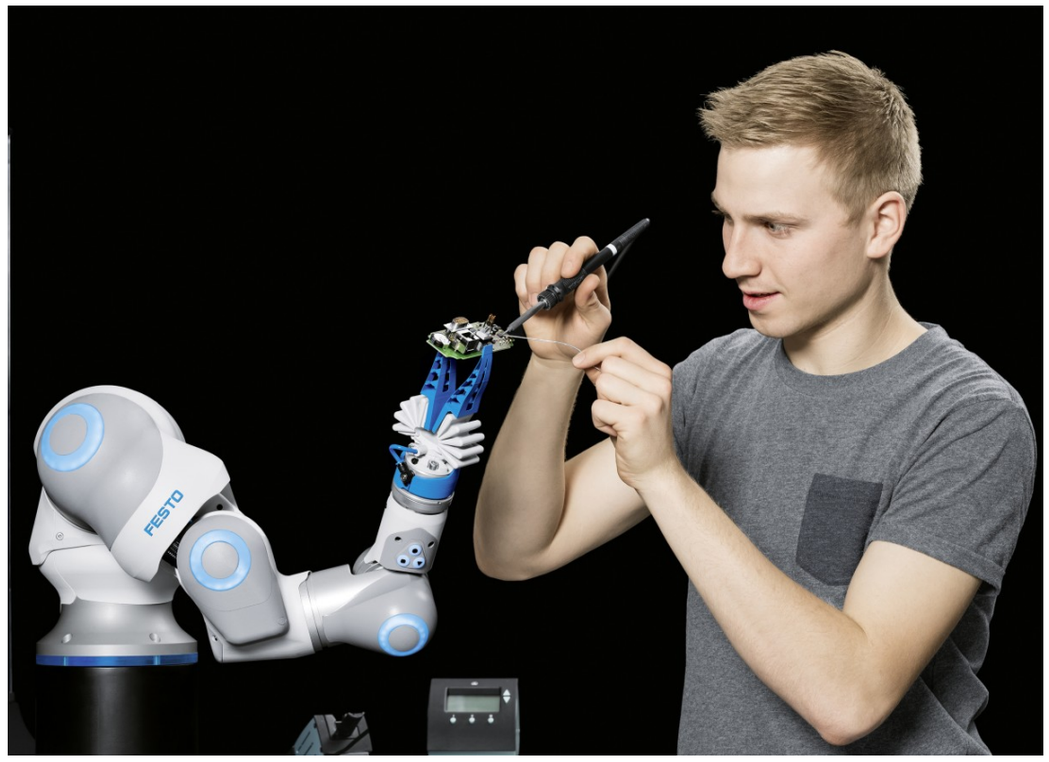


with expert's actions

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

vs.

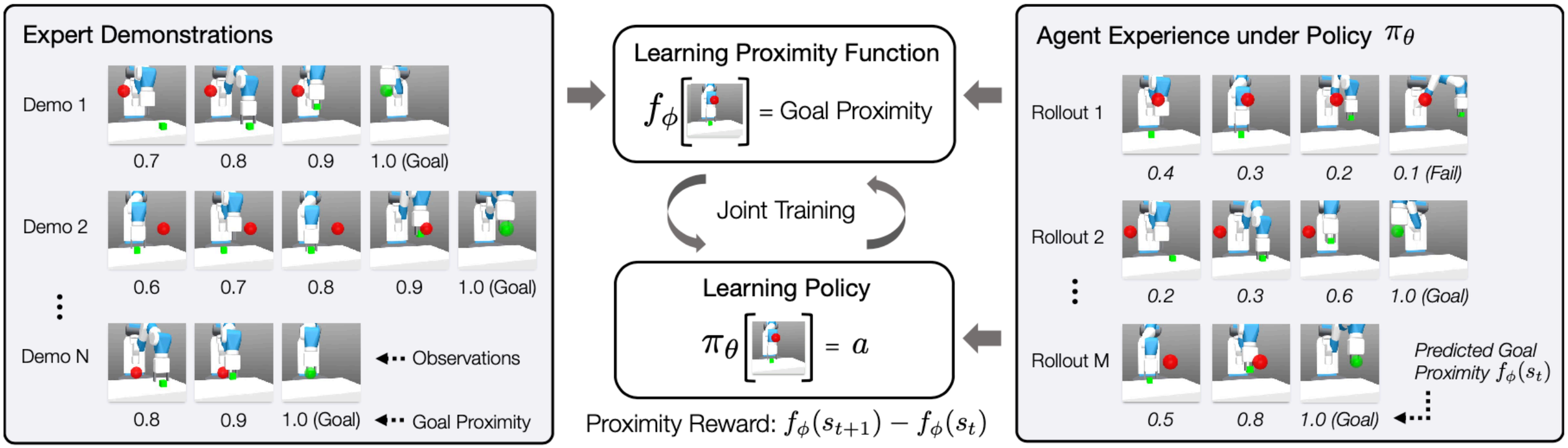
Learning from **Observation**



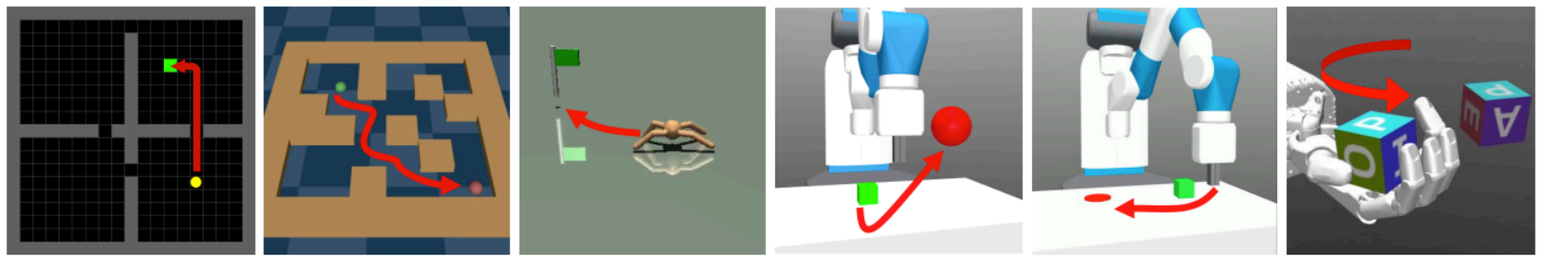
without expert's actions

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

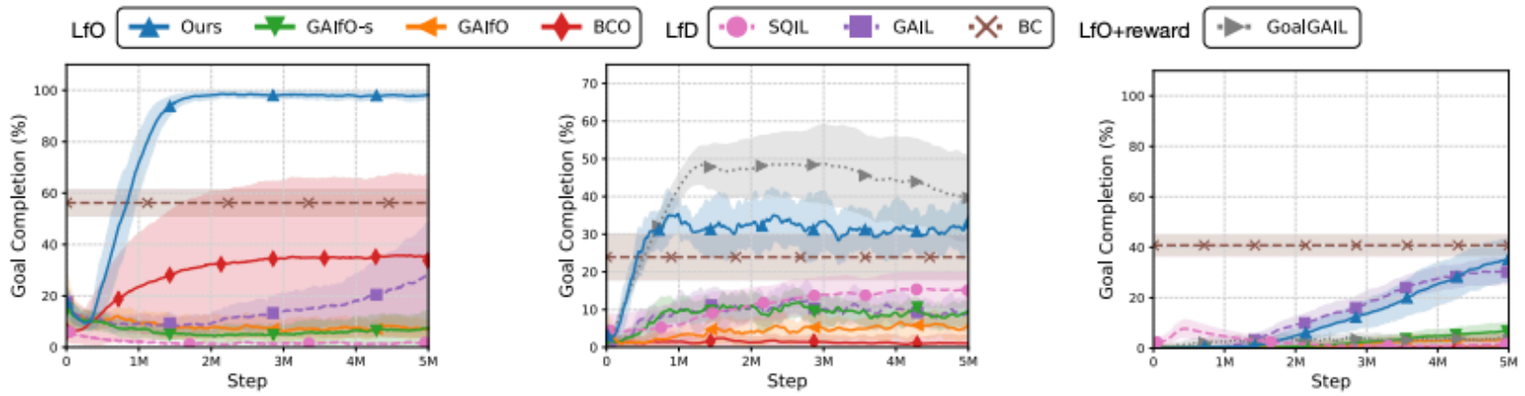
Learning from Observation via Inferring Goal Proximity



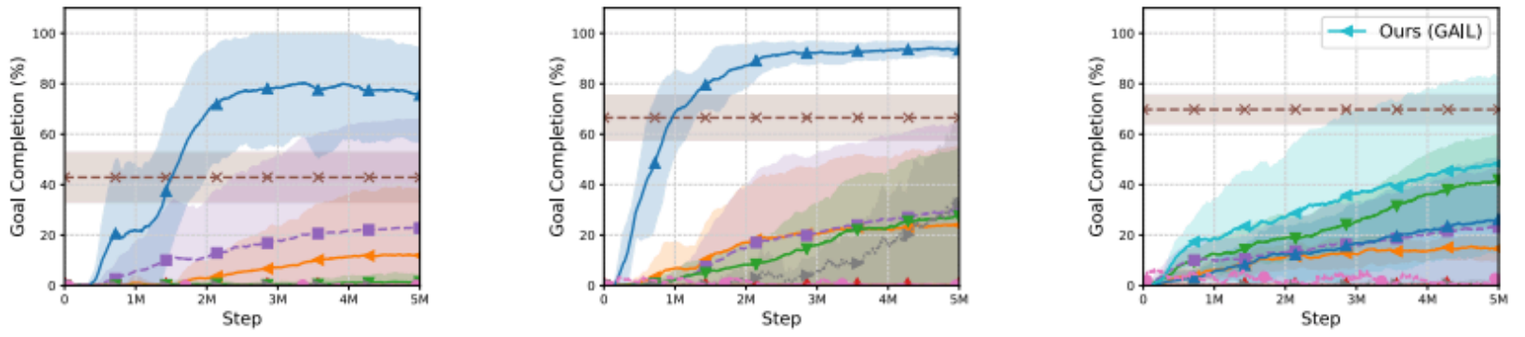
Experiments



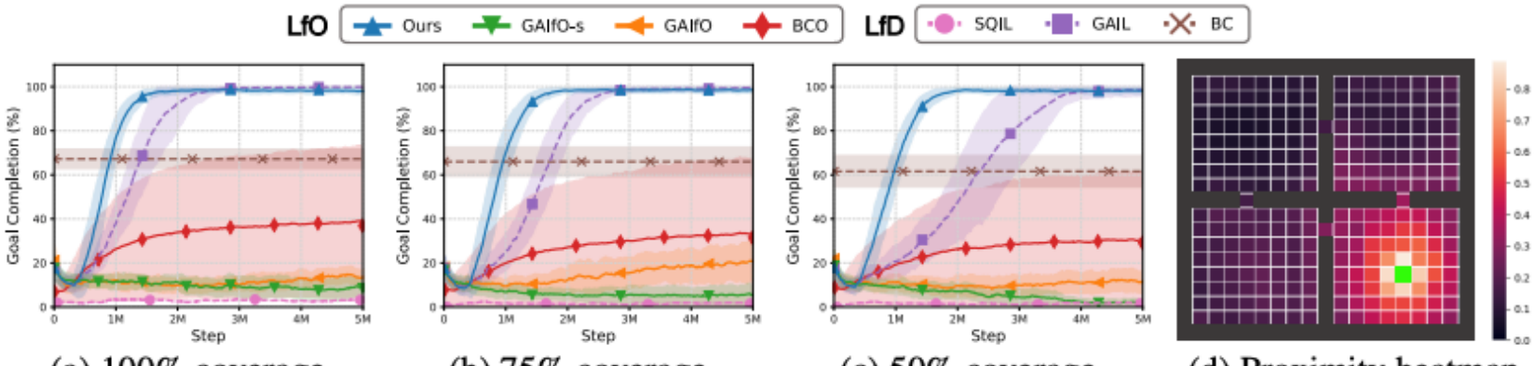
(a) NAVIGATION (b) MAZE2D (c) ANT REACH (d) FETCH PICK (e) FETCH PUSH (f) HAND ROTATE



(a) NAVIGATION 25% (b) MAZE2D 50% (c) ANT REACH 0.05 noise

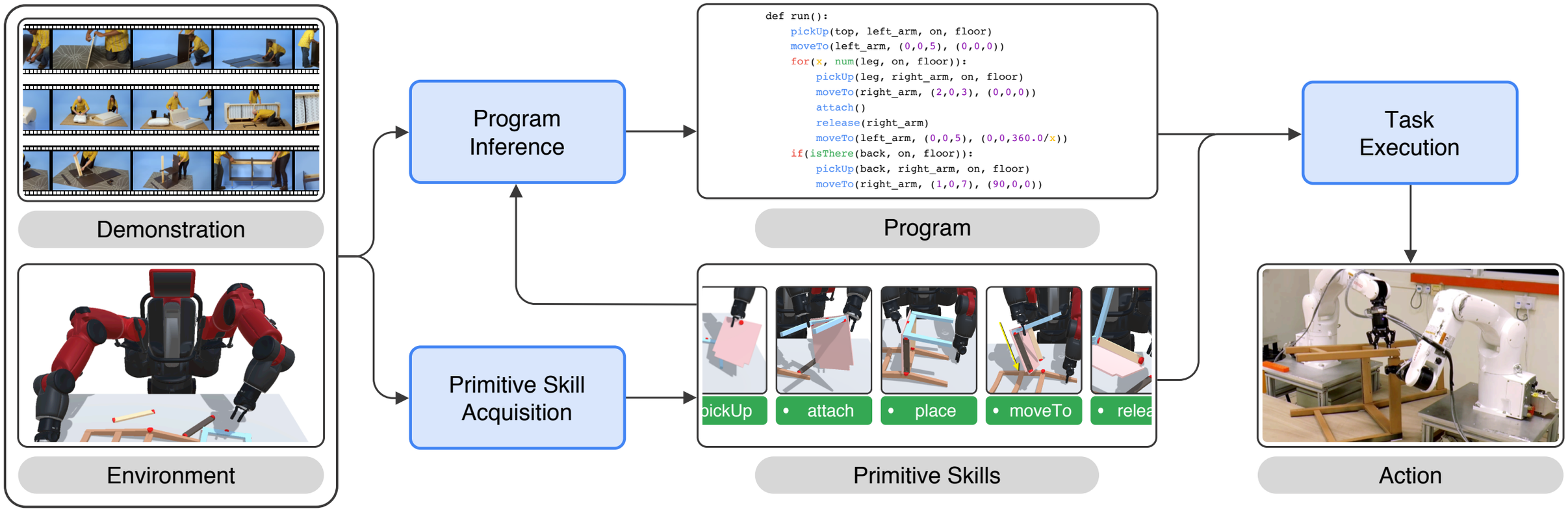


(d) FETCH PICK 1.75x noise (e) FETCH PUSH 1.75x noise (f) HAND ROTATE 0.35 noise



(a) 100% coverage (b) 75% coverage (c) 50% coverage (d) Proximity heatmap

Program-Guided Framework for Interpreting and Acquiring Complex Skills



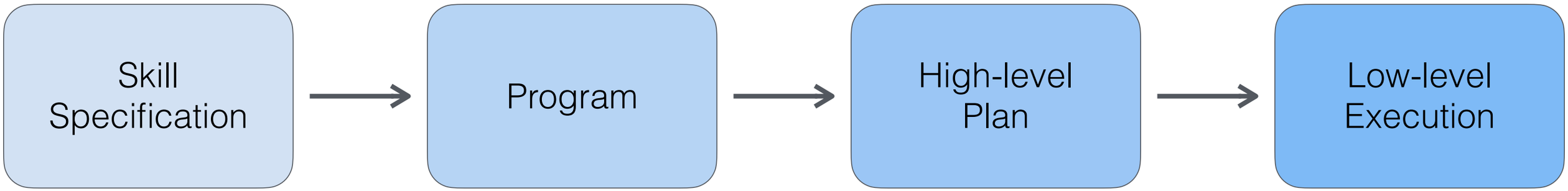
Interpretable

Programmatic / Generalizable

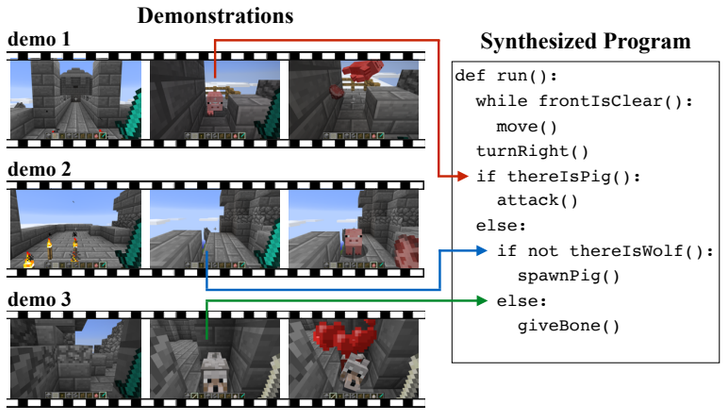
Hierarchical

Modular

Program-Guided Framework for Interpreting and Acquiring Complex Skills



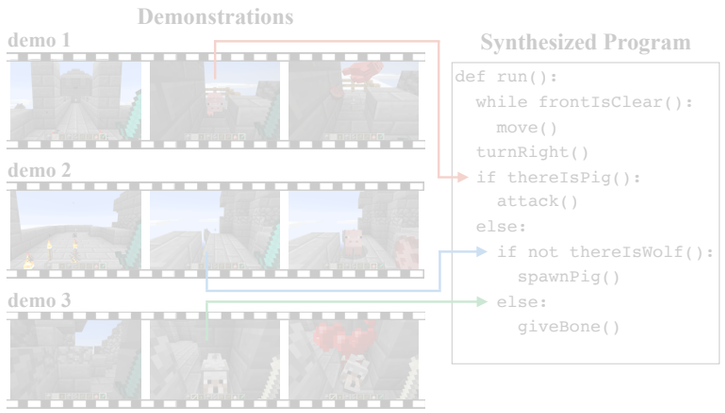
Neural Program Synthesis from Diverse Demonstration Videos



ICML 2018

Skill Specification → Program → High-level Plan → Low-level Execution

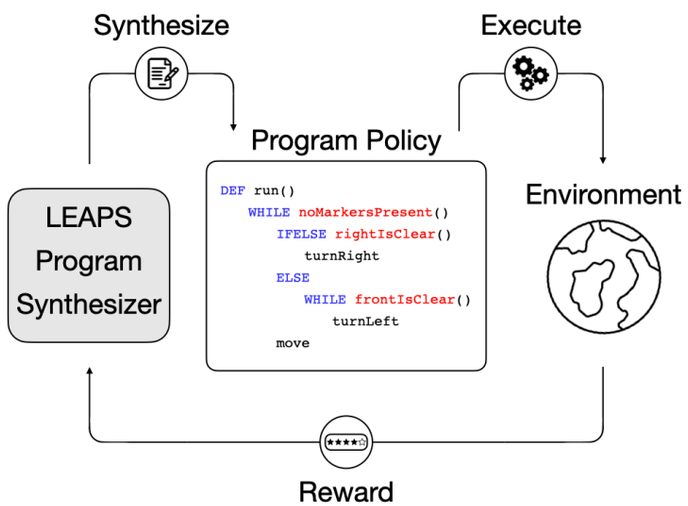
Neural Program Synthesis
from Diverse Demonstration Videos



ICML 2018

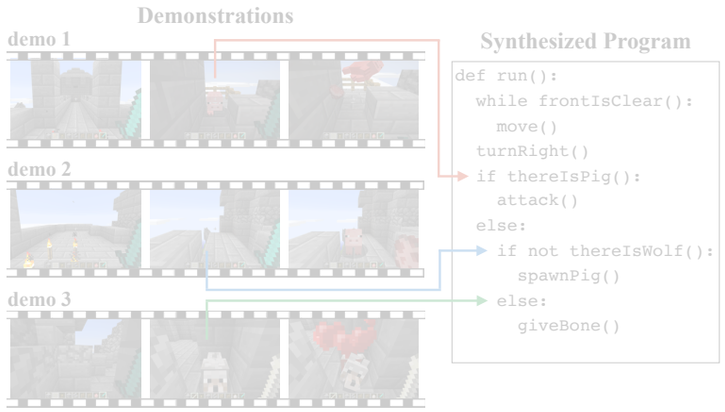
Skill Specification → Program → High-level Plan → Low-level Execution

Learning to Synthesize Programs as
Interpretable and Generalizable Policies



NeurIPS 2021

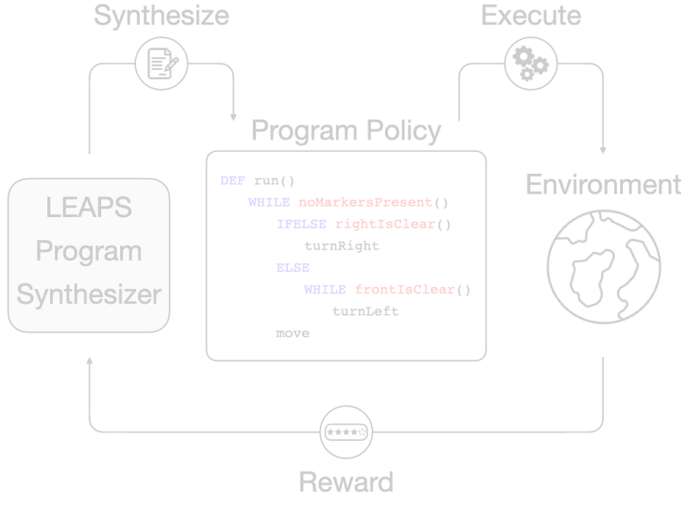
Neural Program Synthesis from Diverse Demonstration Videos



ICML 2018

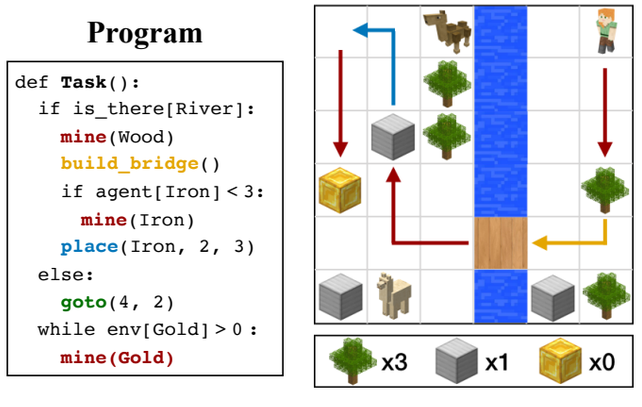
Skill Specification → Program → High-level Plan → Low-level Execution

Learning to Synthesize Programs as Interpretable and Generalizable Policies



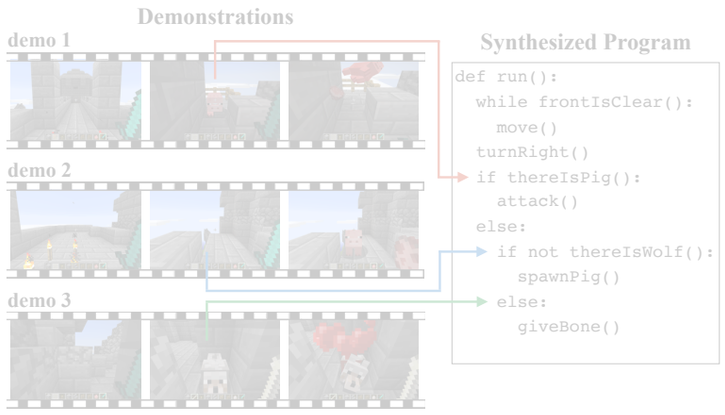
NeurIPS 2021

Program Guided Agent



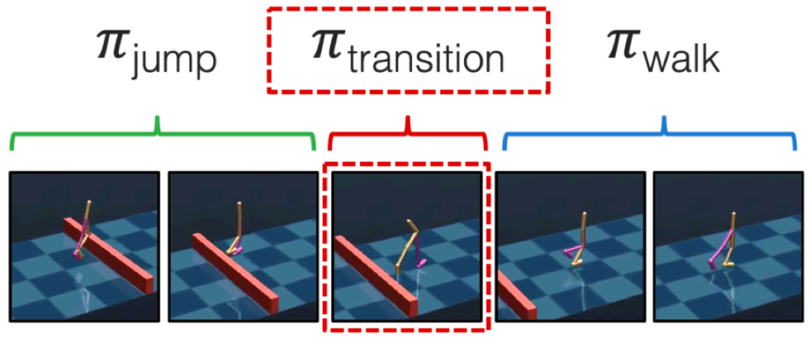
ICLR 2020 (Spotlight)

Neural Program Synthesis from Diverse Demonstration Videos



ICML 2018

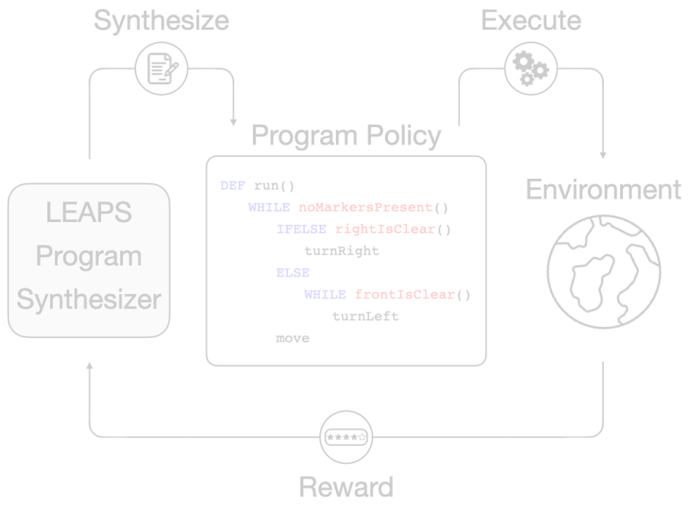
Composing Complex Skills by Learning Transition Policies



ICLR 2019

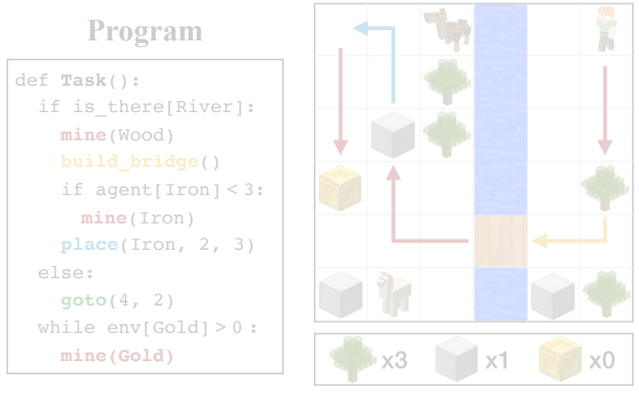
Skill Specification → Program → High-level Plan → Low-level Execution

Learning to Synthesize Programs as Interpretable and Generalizable Policies



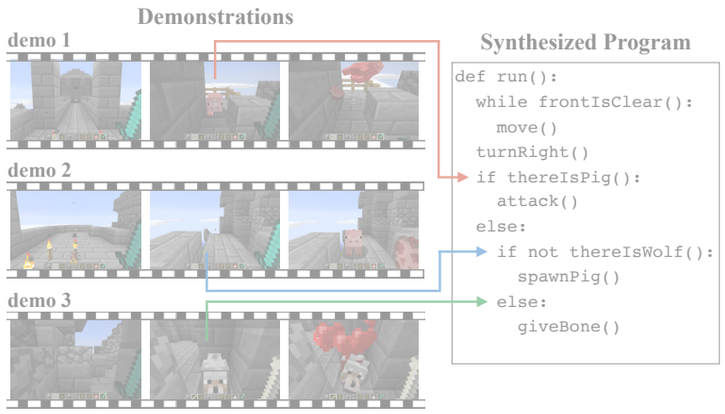
NeurIPS 2021

Program Guided Agent



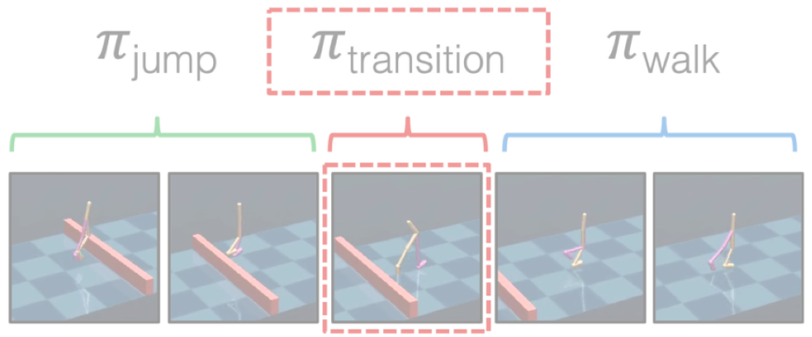
ICLR 2020 (Spotlight)

Neural Program Synthesis from Diverse Demonstration Videos



ICML 2018

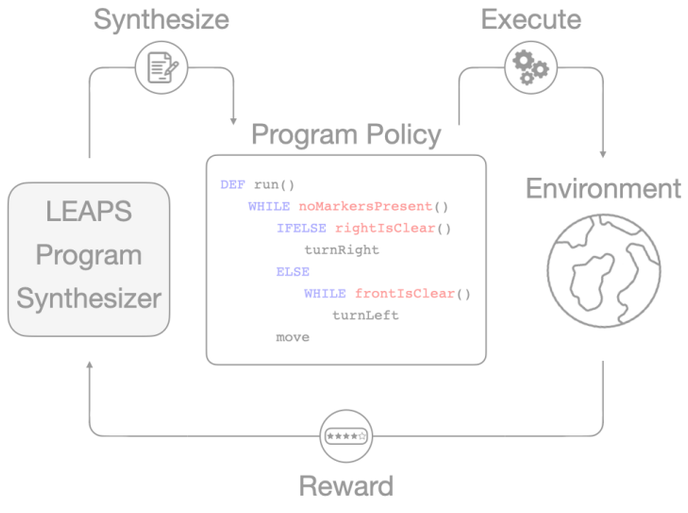
π_{jump} $\pi_{transition}$ π_{walk}



ICLR 2019

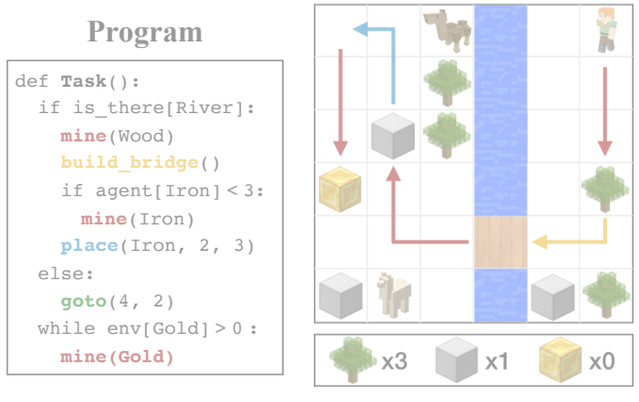
Skill Specification → Program → High-level Plan → Low-level Execution

Learning to Synthesize Programs as Interpretable and Generalizable Policies



NeurIPS 2021

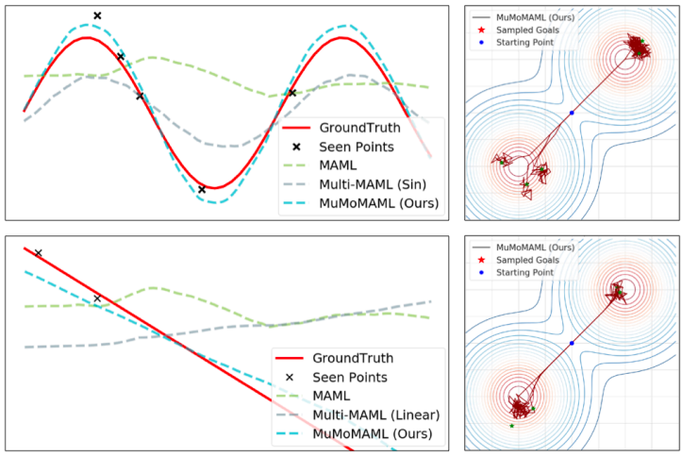
Program Guided Agent



ICLR 2020 (Spotlight)

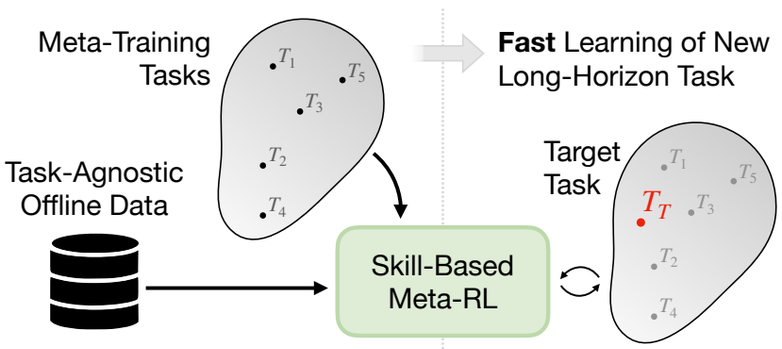
Low-level Execution

Toward Multimodal Model-Agnostic Meta-Learning



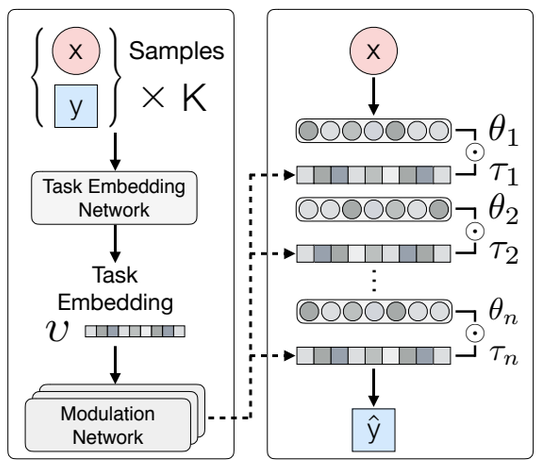
Meta-learning workshop @ NeurIPS 2018

Skill-based Meta-Reinforcement Learning



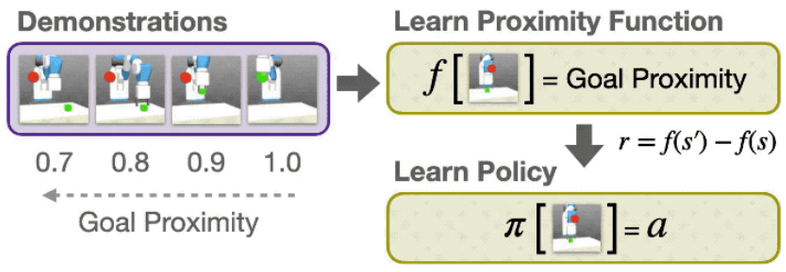
Deep RL workshop @ NeurIPS 2021
Meta-learning workshop @ NeurIPS 2021
submitted to ICLR 2022

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



NeurIPS 2019 (Spotlight)

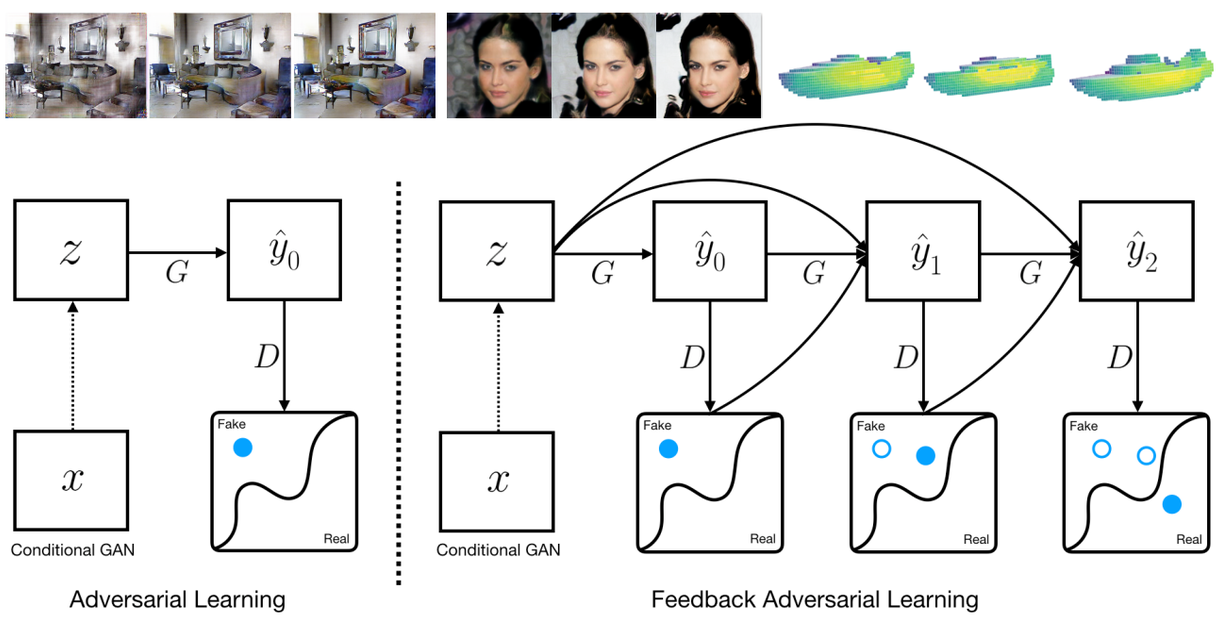
Generalizable Imitation Learning from Observation via Inferring Goal Proximity



NeurIPS 2021

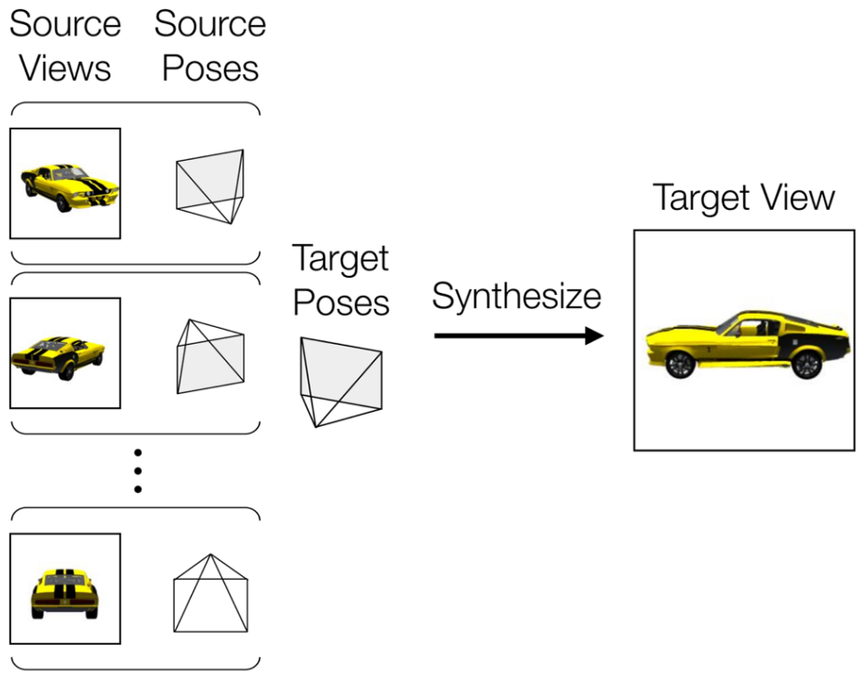
Computer Vision / 3D Vision / Image Synthesis / GAN / Medical Imaging

Feedback Adversarial Learning:
Spatial Feedback for Improving Generative Adversarial Networks



CVPR 2019

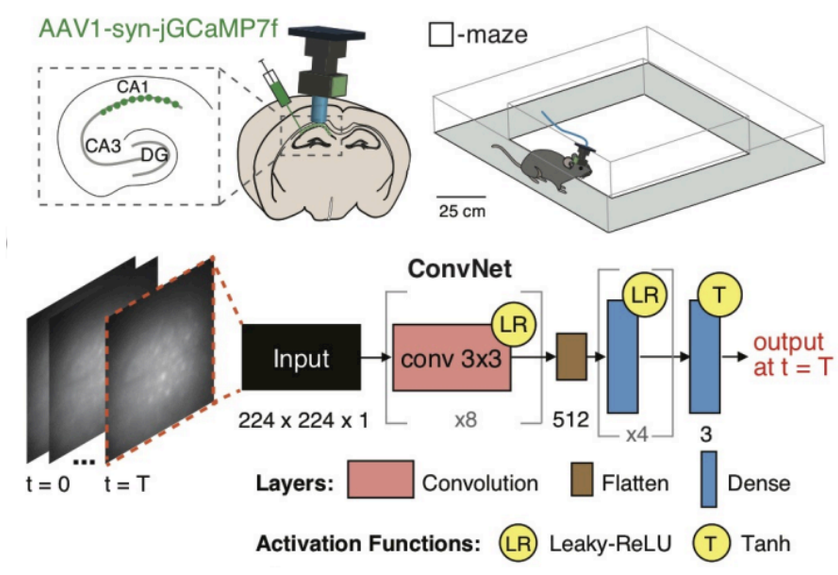
Multi-view to Novel view:
Synthesizing Views with Self-Learned Confidence

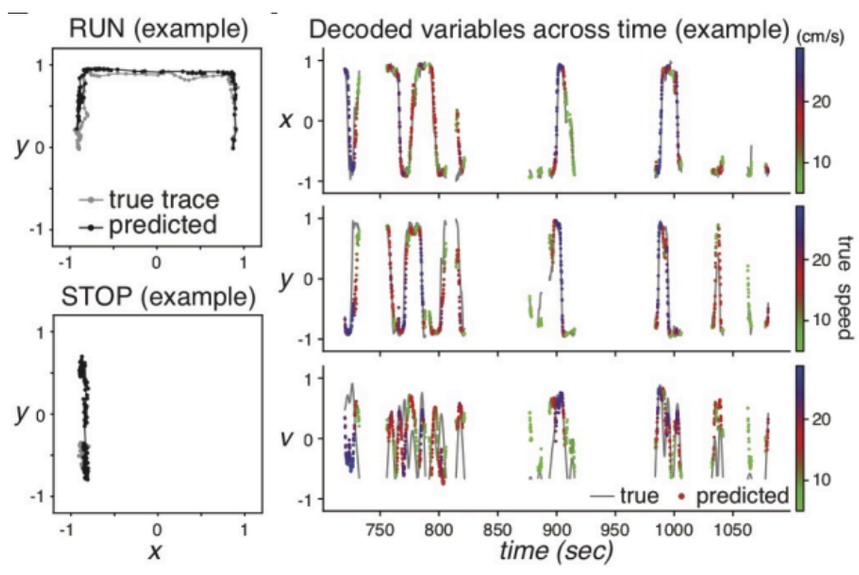


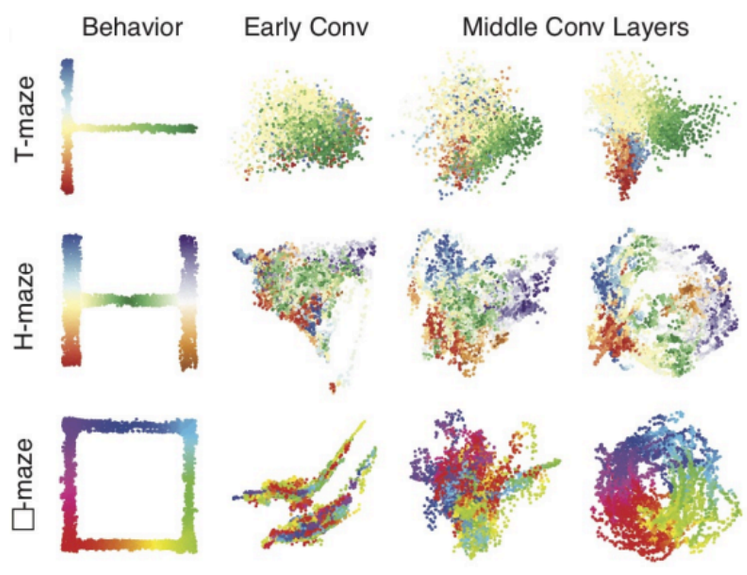
ECCV 2018

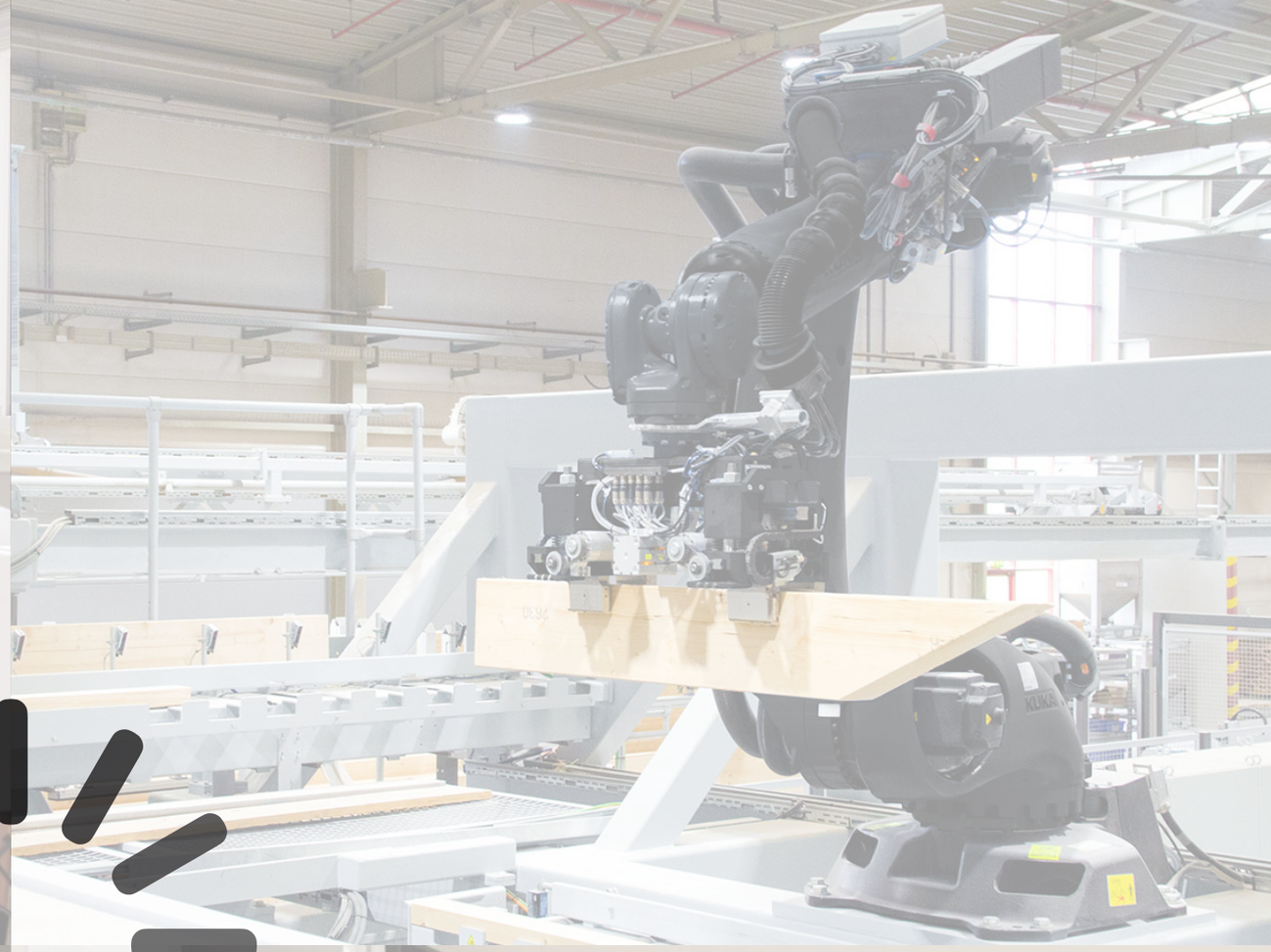
Behavioral Clusters Revealed by End-to-end Decoding from Microendoscopic Imaging

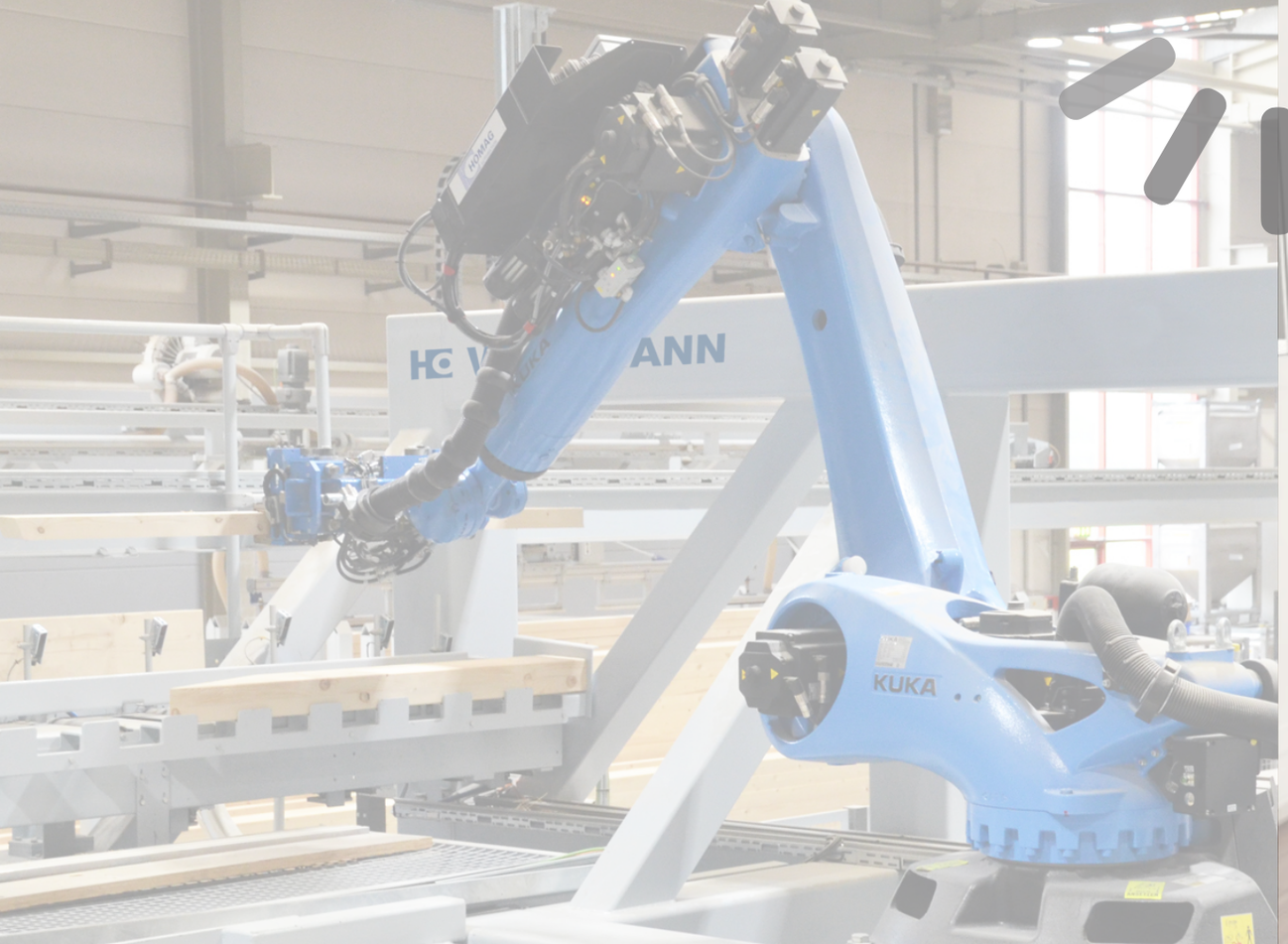
bioRxiv 2021










KUKA

Thank You

Questions?