Towards Grounded Spatio-Temporal Reasoning

Kevin Chih-Yao Ma

Ph.D. student @ Georgia Tech

Outline

- Why do we need grounded spatio-temporal reasoning?
- Related Work & Motivations
- Self-Monitoring and Regretful Navigation Agent
  - Grounding on Vision-and-Language Navigation
- Object Level Fine-Grained Video Understanding
  - Ground human action recognition to object interactions
  - Ground video captioning to object interactions
- Grounded Visual Captioning without human annotations
- Summary

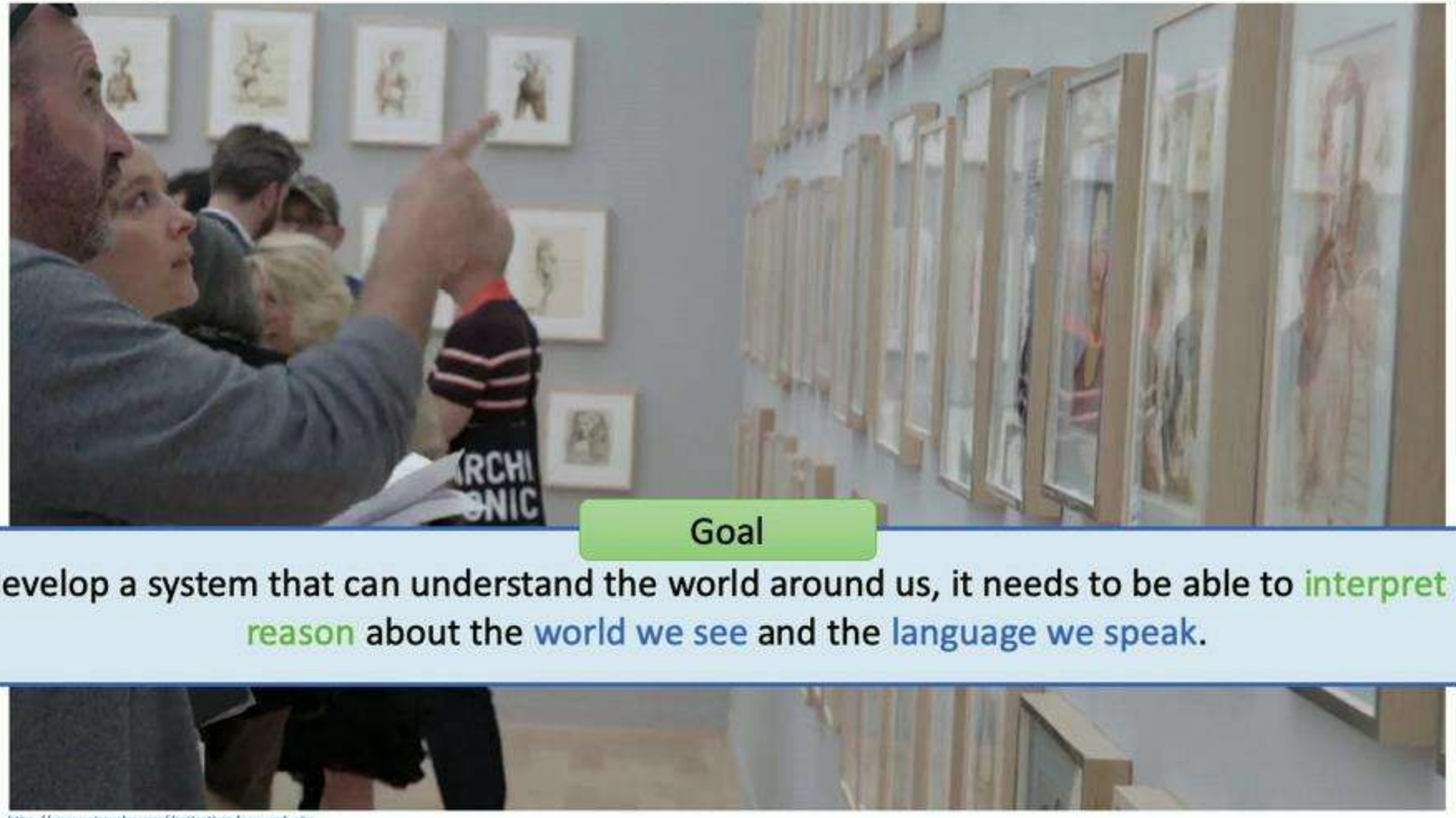
Why Spatio-Temporal Reasoning?

- We live in a structured spatio-temporal space, represented by
  - The image/video we observe — Vision
  - The language we use to describe — Language

https://www.cntraveler.com/destinations/new-york-city

Why Spatio-Temporal Reasoning?

- We live in a structured spatio-temporal space, represented by
  - The image/video we observe — Vision
  - The language we use to describe — Language



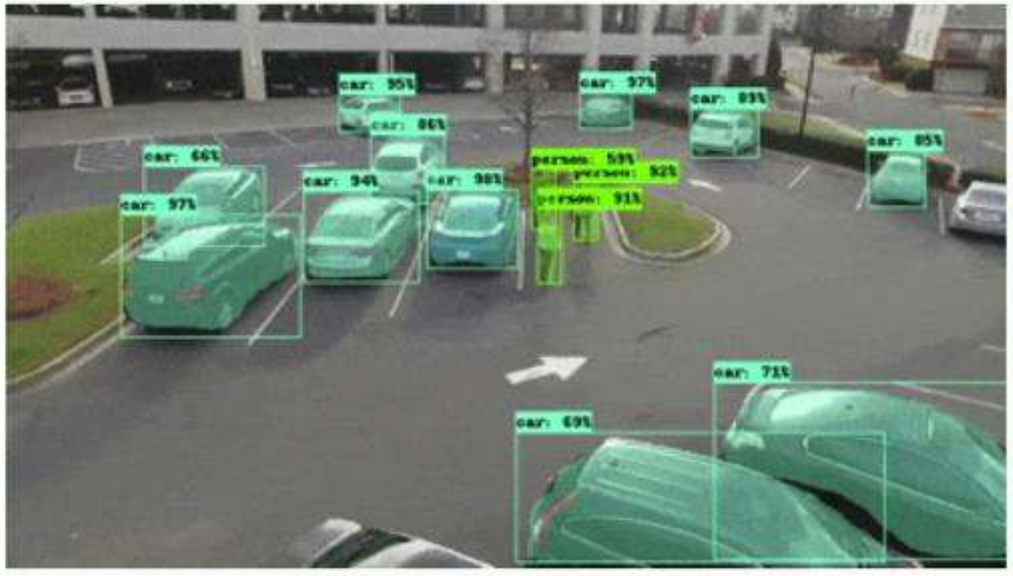
Goal

To develop a system that can understand the world around us, it needs to be able to interpret and reason about the world we see and the language we speak.

https://www.cntraveler.com/destinations/new-york-city

Why Spatio-Temporal Reasoning?

- We live in a structured spatio-temporal space, represented by
  - The image/video we observe — Vision
  - The language we use to describe — Language



[https://www.kdnuggets.com/2018/03/tensorflow-object-detection-pixel-wise-classification.html]
[He et al., ICCV 2017]

[Li et al., CVPR 2017]

Bridging from detection to *reasoning with language*.

Why Grounded Spatio-Temporal Reasoning?

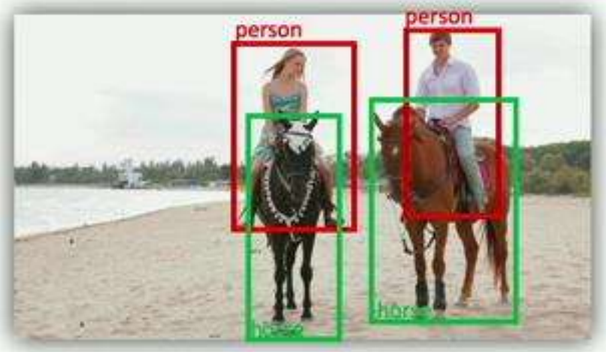
Bridging from detection to *reasoning with language*.

Captioning:

A woman is on a horse.
A man is on a horse.
A man and a woman are on two horses.
A man with a white shirt is riding on a horse.
A man with a white shirt is riding on a horse on a beach.

Why Grounded Spatio-Temporal Reasoning?

Bridging from detection to *reasoning with language*.



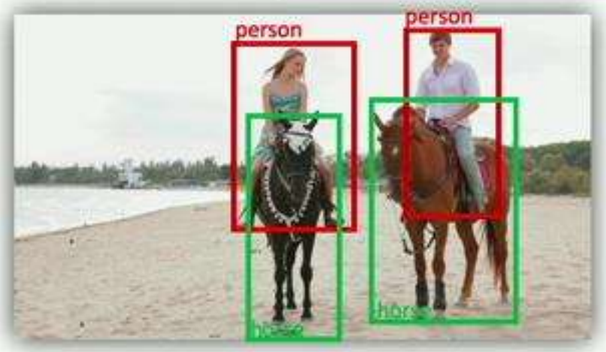
Captioning:

A woman is on a horse.
A man is on a horse.
A man and a woman are on two horses.
A man with a white shirt is riding on a horse.
A man with a white shirt is riding on a horse on a beach.

Language needs to be properly *grounded*.

Why Grounded Spatio-Temporal Reasoning?

Bridging from detection to *reasoning with language*.



Captioning:

A woman is on a horse.
A man is on a horse.
A man and a woman are on two horses.
A man with a white shirt is riding on a horse.
A man with a white shirt is riding on a horse on a beach.

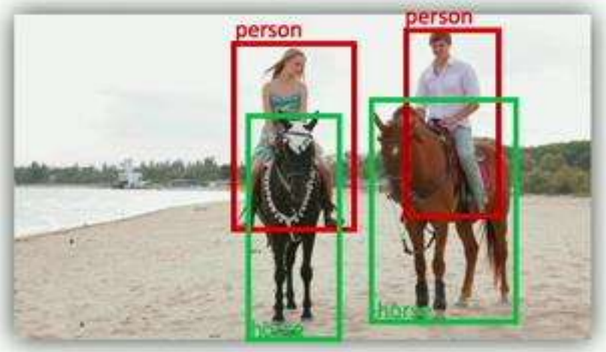
Language needs to be properly *grounded*.

Vision-and-Language Navigation

Leave the bedroom, and enter the kitchen. Walk forward, and take a left at the couch. Stop in front of the window.

Why Grounded Spatio-Temporal Reasoning?

Bridging from detection to *reasoning with language*.



Captioning:

A woman is on a horse.
A man is on a horse.
A man and a woman are on two horses.
A man with a white shirt is riding on a horse.
A man with a white shirt is riding on a horse on a beach.

Language needs to be properly *grounded*.

Vision-and-Language Navigation

Leave the bedroom, and enter the kitchen. Walk forward, and take a left at the couch. Stop in front of the window.

Action Recognition

Outline

- Why do we need grounded spatio-temporal reasoning?
- Related Work & Motivations
- Self-Monitoring and Regretful Navigation Agent
  - Grounding on Vision-and-Language Navigation
- Object Level Fine-Grained Video Understanding
  - Ground human action recognition to object interactions
  - Ground video captioning to object interactions
- Grounded Visual Captioning without human annotations
- Summary

Related Work and Motivations

Vision-and-Language Navigation

- Anderson et al., CVPR 2018
- Wang et al., ECCV 2018
- Fried et al., NeurIPS 2018

seq-to-seq

soft-attention

RL

Exit the bedroom and go towards the table. Go to the stairs on the left of the couch. Wait on the third step.

Related Work and Motivations

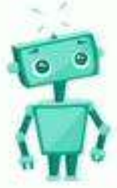
Vision-and-Language Navigation

- Anderson et al., CVPR 2018
- Wang et al., ECCV 2018
- Fried et al., NeurIPS 2018

seq-to-seq

soft-attention

RL



Exit the bedroom and go towards the table. Go to the stairs on the left of the couch. Wait on the third step.

by *ent423* , *ent261* correspondent updated 9:49 pm et , thu march 19 , 2015 (*ent261*) a *ent114* was killed in a parachute accident in *ent45* , *ent85* , near *ent312* , a *ent119* official told *ent261* on wednesday . he was identified thursday as special warfare operator 3rd class *ent23* , 29 , of *ent187* , *ent265* . `` *ent23* distinguished himself consistently throughout his career . he was the epitome of the quiet professional in all facets of his life , and he leaves an inspiring legacy of natural tenacity and focused

. . .

[Hermann et al., NeurIPS 2015]

Related Work and Motivations

Vision-and-Language Navigation

- Anderson et al., CVPR 2018
- Wang et al., ECCV 2018
- Fried et al., NeurIPS 2018

seq-to-seq

soft-attention

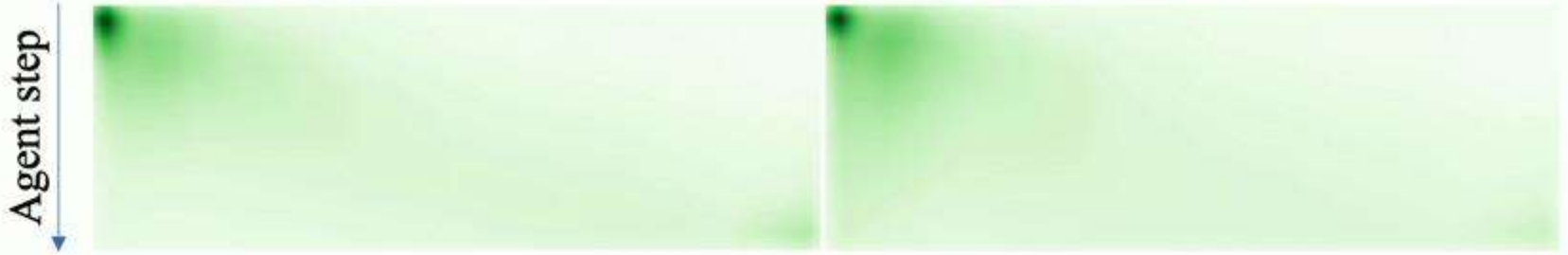
RL

Exit the bedroom and go towards the table. Go to the stairs on the left of the couch. Wait on the third step.

Incorrect grounding from previous work [1]

validation seen

validation unseen



instruction: $x_1, x_2, \ldots, x_L$

[1] Speaker-Follower (Fried et al., NeurIPS 2018)

Related Work and Motivations

Vision-and-Language Navigation

- Anderson et al., CVPR 2018
- Wang et al., ECCV 2018
- Fried et al., NeurIPS 2018

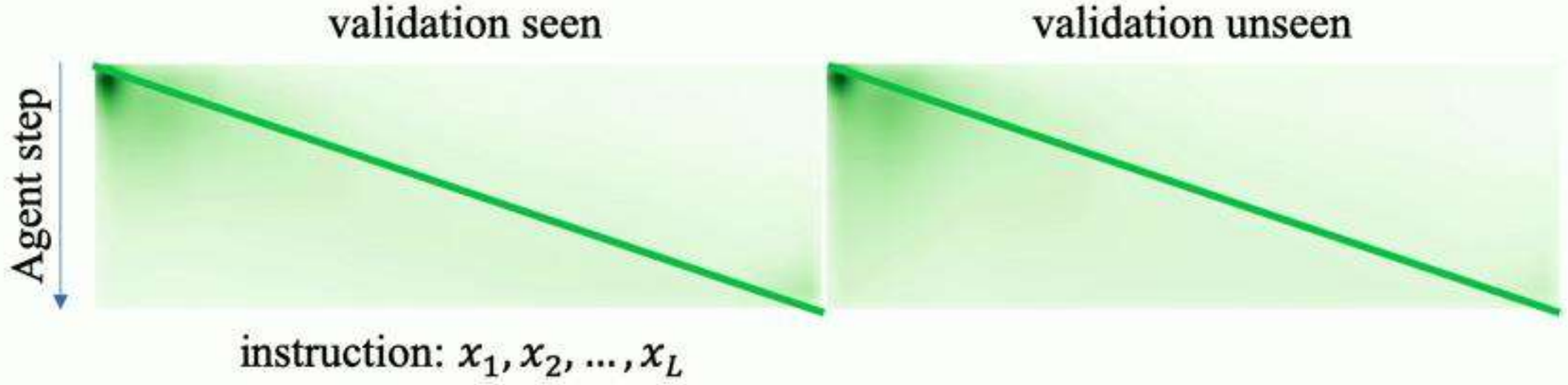
seq-to-seq

soft-attention

RL

Exit the bedroom and go towards the table. Go to the stairs on the left of the couch. Wait on the third step.

Incorrect grounding from previous work [1]



[1] Speaker-Follower (Fried et al., NeurIPS 2018)

Related Work and Motivations

Grounding issue

Vision-and-Language Navigation

- Anderson et al., CVPR 2018
- Wang et al., ECCV 2018
- Fried et al., NeurIPS 2018

seq-to-seq

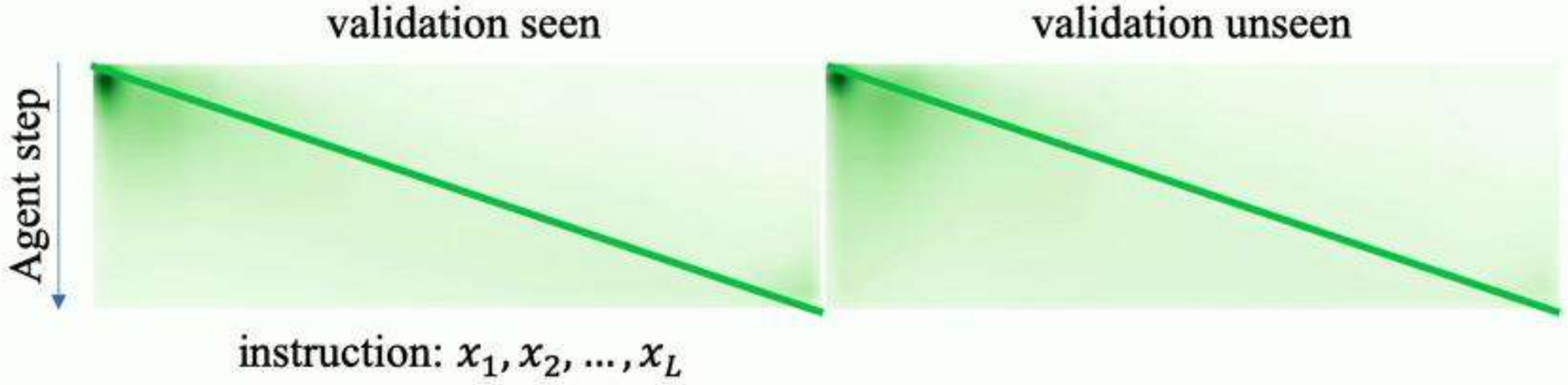
soft-attention

RL

⚠ Not properly grounded

Exit the bedroom and go towards the table. Go to the stairs on the left of the couch. Wait on the third step.

Incorrect grounding from previous work [1]



[1] Speaker-Follower (Fried et al., NeurIPS 2018)

Related Work and Motivations

Grounding issue

Vision-and-Language Navigation

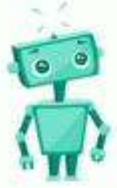
- Anderson et al., CVPR 2018
- Wang et al., ECCV 2018
- Fried et al., NeurIPS 2018

seq-to-seq

soft-attention

RL

⚠ Not properly grounded



Exit the bedroom and go towards the table. Go to the stairs on the left of the couch. Wait on the third step.

Goal

Enforce grounding on the instruction for navigation agent

Related Work and Motivations

Action Recognition

- Simonyan et al., NeurIPS 2014
- Feichtenhofer et al., CVPR 2016
- Sigurdsson et al., CVPR 2016
- Girdhar et al., CVPR 2017
- Carreira et al., CVPR 2017
- Qiu et al., ICCV 2017

LSTM

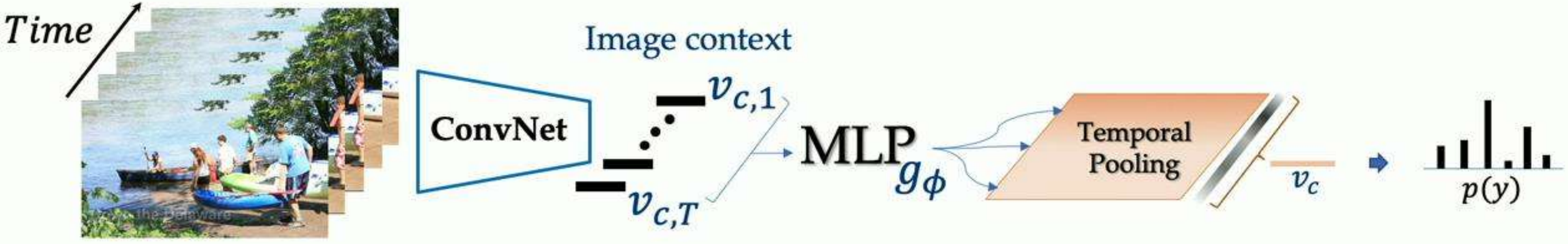
1D or 3D Conv

CRF

VLAD

optical flow

Existing approaches rely on frame-level representation



Related Work and Motivations

Action Recognition

- Simonyan et al., NeurIPS 2014
- Feichtenhofer et al., CVPR 2016
- Sigurdsson et al., CVPR 2016
- Girdhar et al., CVPR 2017
- Carreira et al., CVPR 2017
- Qiu et al., ICCV 2017

LSTM

1D or 3D Conv

CRF

VLAD

optical flow

Human action recognition involves complex interactions between objects

Skiing

≠

Snowboarding

Related Work and Motivations

Action Recognition

- Simonyan et al., NeurIPS 2014
- Feichtenhofer et al., CVPR 2016
- Sigurdsson et al., CVPR 2016
- Girdhar et al., CVPR 2017
- Carreira et al., CVPR 2017
- Qiu et al., ICCV 2017

LSTM

1D or 3D Conv

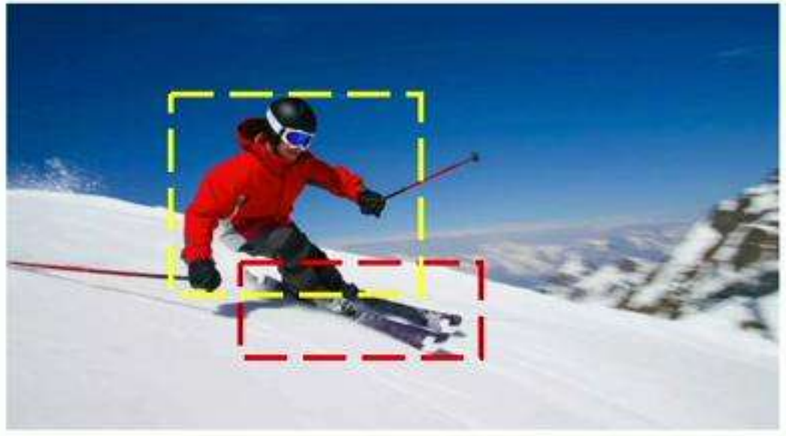
CRF

VLAD

optical flow

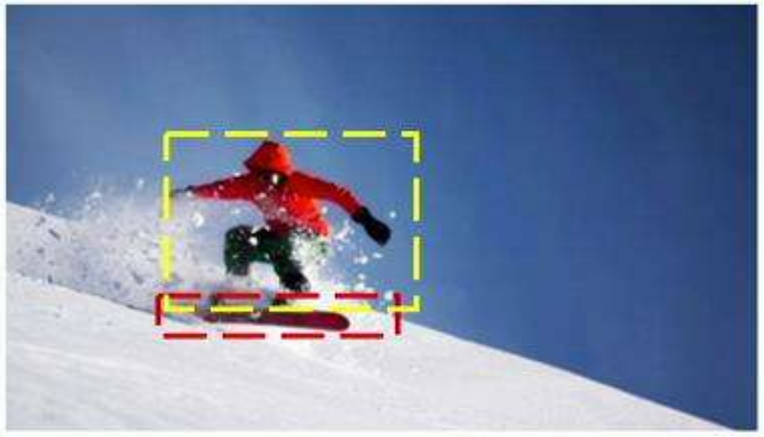
Human action recognition involves complex interactions between objects

Skiing



≠

Snowboarding



[Person] riding [ski]

[Person] riding [snowboard]

Related Work and Motivations

Grounding issue

Action Recognition

- Simonyan et al., NeurIPS 2014
- Feichtenhofer et al., CVPR 2016
- Sigurdsson et al., CVPR 2016
- Girdhar et al., CVPR 2017
- Carreira et al., CVPR 2017
- Qiu et al., ICCV 2017

LSTM

1D or 3D Conv

CRF

VLAD

optical flow

⚠ No elements to ground

Human action recognition involves complex interactions between objects

Skiing

≠

Snowboarding

[Person] riding [ski]

[Person] riding [snowboard]

Related Work and Motivations

Grounding issue

Action Recognition

- Simonyan et al., NeurIPS 2014
- Feichtenhofer et al., CVPR 2016
- Sigurdsson et al., CVPR 2016
- Girdhar et al., CVPR 2017
- Carreira et al., CVPR 2017
- Qiu et al., ICCV 2017

LSTM

1D or 3D Conv

CRF

VLAD

optical flow

⚠ No elements to ground

Goal

Object-level video understanding

Ground human actions to object interactions

Related Work and Motivations

Video Captioning

- Venugopalan et al., ACL 2014
- Yao et al., ICCV 2015
- Yu et al., CVPR 2016
- Gan et al., CVPR 2017
- Pan et al., CVPR 2017
- Shen et al., CVPR 2017
- Lu et al., CVPR 2018
- Zhou et al., CVPR 2019

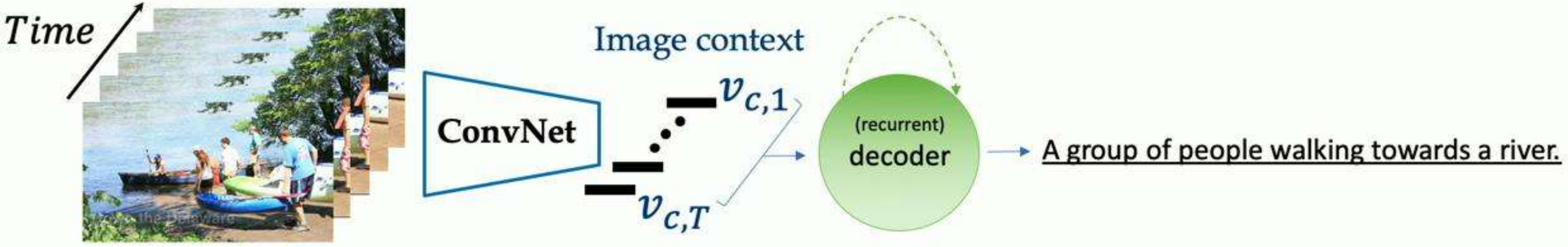
seq-to-seq

soft-attention

semantic attribute

RL

Existing approaches rely on frame-level representation



Related Work and Motivations

Grounding issue

Video Captioning

- Venugopalan et al., ACL 2014
- Yao et al., ICCV 2015
- Yu et al., CVPR 2016
- Gan et al., CVPR 2017
- Pan et al., CVPR 2017
- Shen et al., CVPR 2017
- Lu et al., CVPR 2018
- Zhou et al., CVPR 2019

seq-to-seq

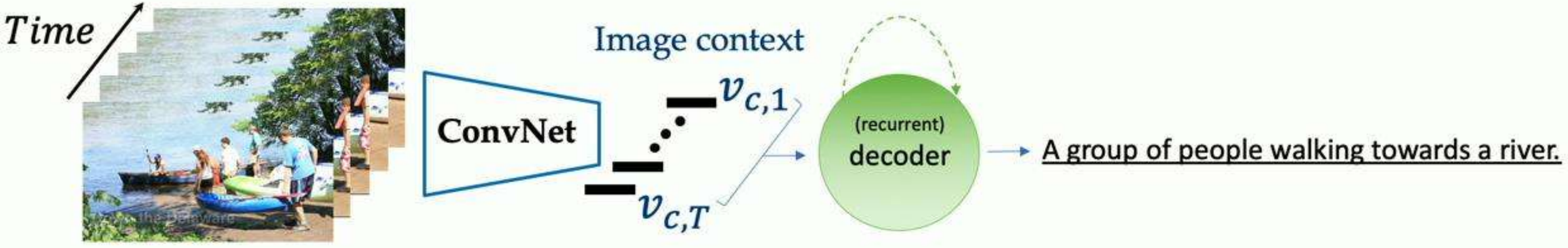
soft-attention

semantic attribute

RL

⚠ No elements to ground

Existing approaches rely on frame-level representation



Related Work and Motivations

Grounding issue

Visual Captioning

- Venugopalan et al., ACL 2014
- Yao et al., ICCV 2015
- Yu et al., CVPR 2016
- Gan et al., CVPR 2017
- Pan et al., CVPR 2017
- Shen et al., CVPR 2017
- Lu et al., CVPR 2018
- Zhou et al., CVPR 2019

seq-to-seq

soft-attention

semantic attribute

RL

⚠ No elements to ground

Goal

Object-level video understanding

Ground visual descriptions to object interactions

Related Work and Motivations

Grounding issue

Visual Captioning

- Venugopalan et al., ACL 2014
- Yao et al., ICCV 2015
- Yu et al., CVPR 2016
- Gan et al., CVPR 2017
- Pan et al., CVPR 2017
- Shen et al., CVPR 2017
- Lu et al., CVPR 2018
- Zhou et al., CVPR 2019

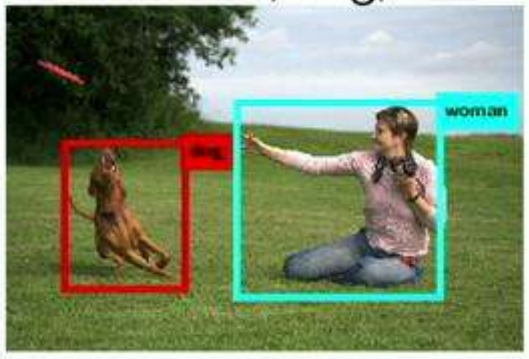
seq-to-seq

soft-attention

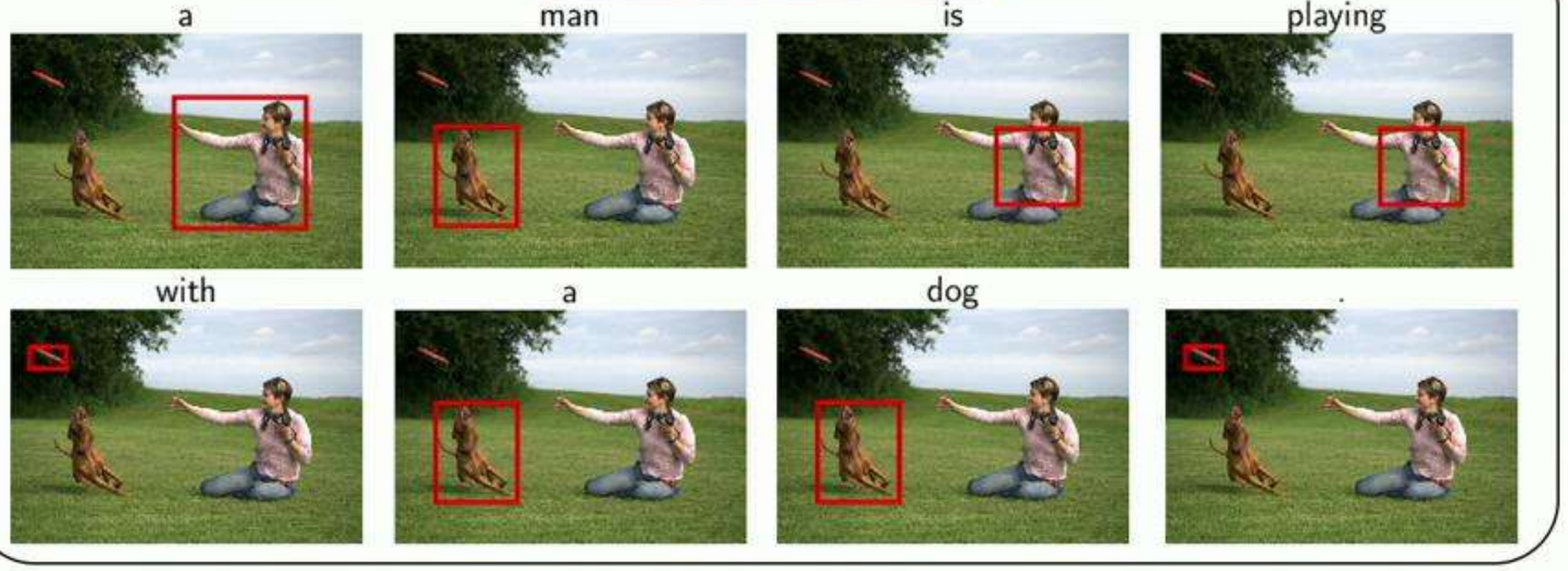
semantic attribute

RL

woman, dog,



Not properly grounded



Related Work and Motivations

Grounding issue

Visual Captioning

- Venugopalan et al., ACL 2014
- Yao et al., ICCV 2015
- Yu et al., CVPR 2016
- Gan et al., CVPR 2017
- Pan et al., CVPR 2017
- Shen et al., CVPR 2017
- Lu et al., CVPR 2018
- Zhou et al., CVPR 2019

seq-to-seq

soft-attention

semantic attribute

RL

⚠ Grounding relies on supervision

Existing approaches improve grounding with ground-truth annotations

A man in a striped shirt is playing the piano on the street while people watch him. [1]

[1] Zhou et al., CVPR 2019

Related Work and Motivations

Grounding issue

Visual Captioning

- Venugopalan et al., ACL 2014
- Yao et al., ICCV 2015
- Yu et al., CVPR 2016
- Gan et al., CVPR 2017
- Pan et al., CVPR 2017
- Shen et al., CVPR 2017
- Lu et al., CVPR 2018
- Zhou et al., CVPR 2019

seq-to-seq

soft-attention

semantic attribute

RL

⚠ Grounding relies on supervision

Goal

Enforce grounding of the captioning model without ground-truth grounding annotations.

Related Work and Motivations

Grounding issue

Vision-and-Language Navigation

- Anderson et al., CVPR 2018
- Wang et al., ECCV 2018
- Fried et al., NeurIPS 2018

seq-to-seq

soft-attention

RL

⚠ Not properly grounded

Action Recognition

- Simonyan et al., NeurIPS 2014
- Feichtenhofer et al., CVPR 2016
- Sigurdsson et al., CVPR 2016
- Girdhar et al., CVPR 2017
- Carreira et al., CVPR 2017
- Qiu et al., ICCV 2017

LSTM

1D or 3D Conv

CRF

VLAD

optical flow

⚠ No elements to ground

Visual Captioning

- Venugopalan et al., ACL 2014
- Yao et al., ICCV 2015
- Yu et al., CVPR 2016
- Gan et al., CVPR 2017
- Pan et al., CVPR 2017
- Shen et al., CVPR 2017
- Lu et al., CVPR 2018
- Zhou et al., CVPR 2019

seq-to-seq

soft-attention

semantic attribute

RL

⚠ No elements to ground
Grounding relies on supervision

Outline

- Why do we need grounded spatio-temporal reasoning?
- Related Work & Motivations
- Self-Monitoring and Regretful Navigation Agent
  - Grounding on Vision-and-Language Navigation
- Object Level Fine-Grained Video Understanding
  - Ground human action recognition to object interactions
  - Ground video captioning to object interactions
- Grounded Visual Captioning without human annotations
- Summary

Vision-and-Language Navigation (VLN)

Exit the bedroom and go towards the table. Go to the stairs on the left of the couch. Wait on the third step.

Matterport3D

Panoramic camera

Vision-and-Language Navigation (VLN)

Exit the bedroom and go towards the table. Go to the stairs on the left of the couch. Wait on the third step.

Matterport3D

Panoramic camera

Challenges:

Association between language and images

Exit *Bedroom* 

Vision-and-Language Navigation (VLN)

Exit the bedroom and go towards the table. Go to the stairs on the left of the couch. Wait on the third step.

Matterport3D

Panoramic camera

Challenges:

Association between language and images

Exit *Bedroom*

No map or GPS

Vision-and-Language Navigation (VLN)

Georgia Tech

Exit the bedroom and go towards the table. Go to the stairs on the left of the couch. Wait on the third step.

Matterport3D

Panoramic camera

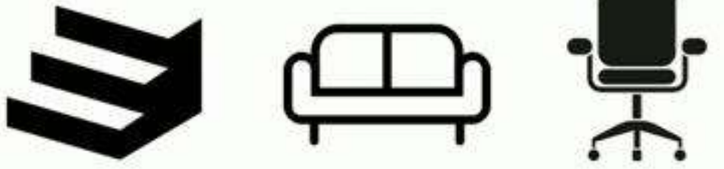
Challenges:

Association between language and images

Exit *Bedroom* 

No map or GPS

No explicit object as the goal



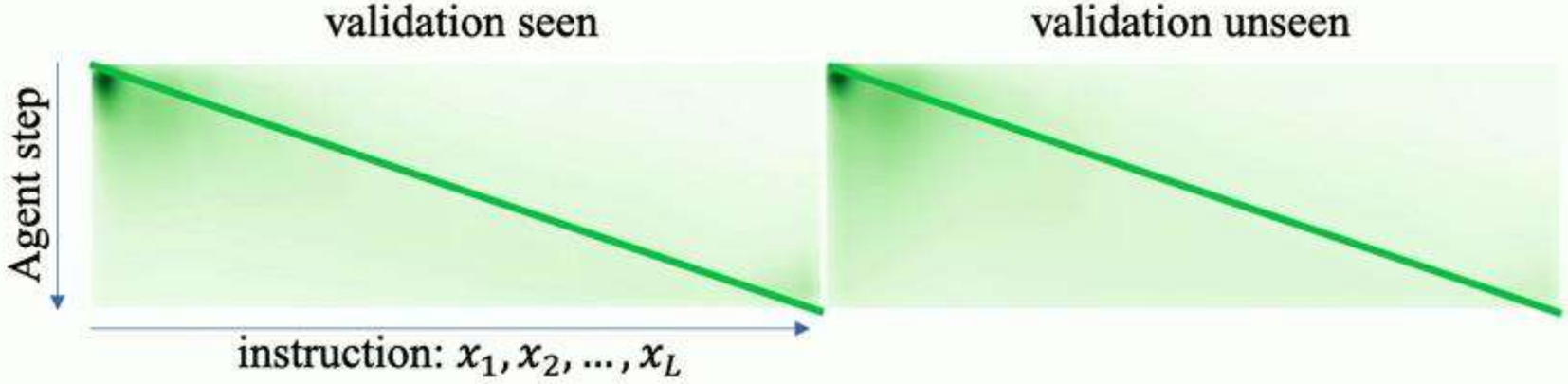
Correctness of Grounding on Instructions

Exit the bedroom and go towards the table. Go to the stairs on the left of the couch. Wait on the third step.

Correctness of Grounding on Instructions

Exit the bedroom and go towards the table. Go to the stairs on the left of the couch. Wait on the third step.

Incorrect grounding from previous work [1]

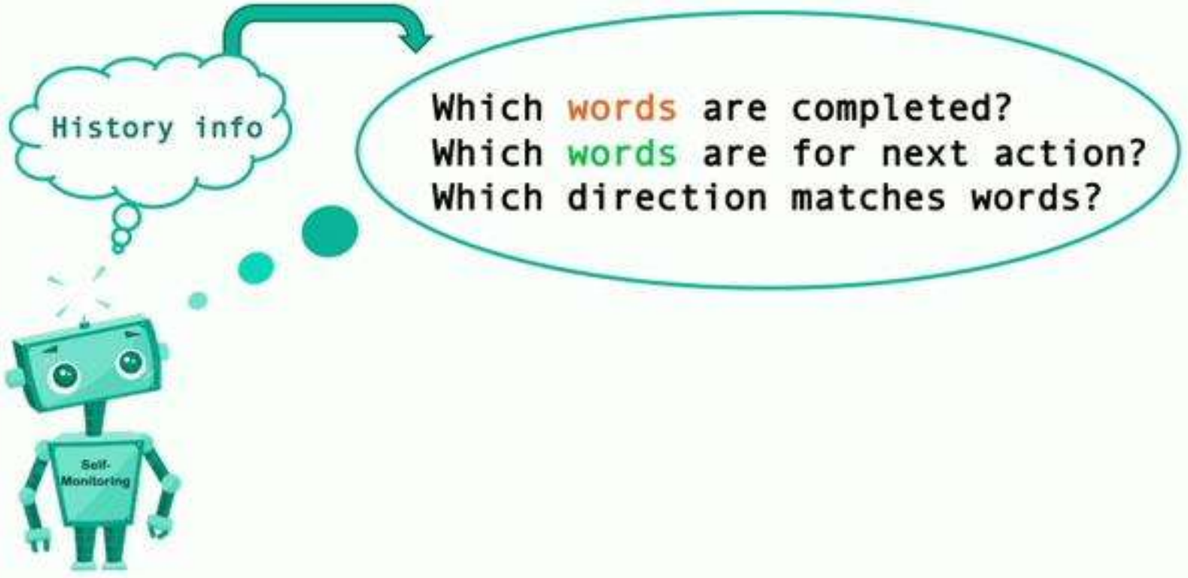


[1] Speaker-Follower (Fried et al., NeurIPS 2018)

Self-Monitoring Navigation Agent

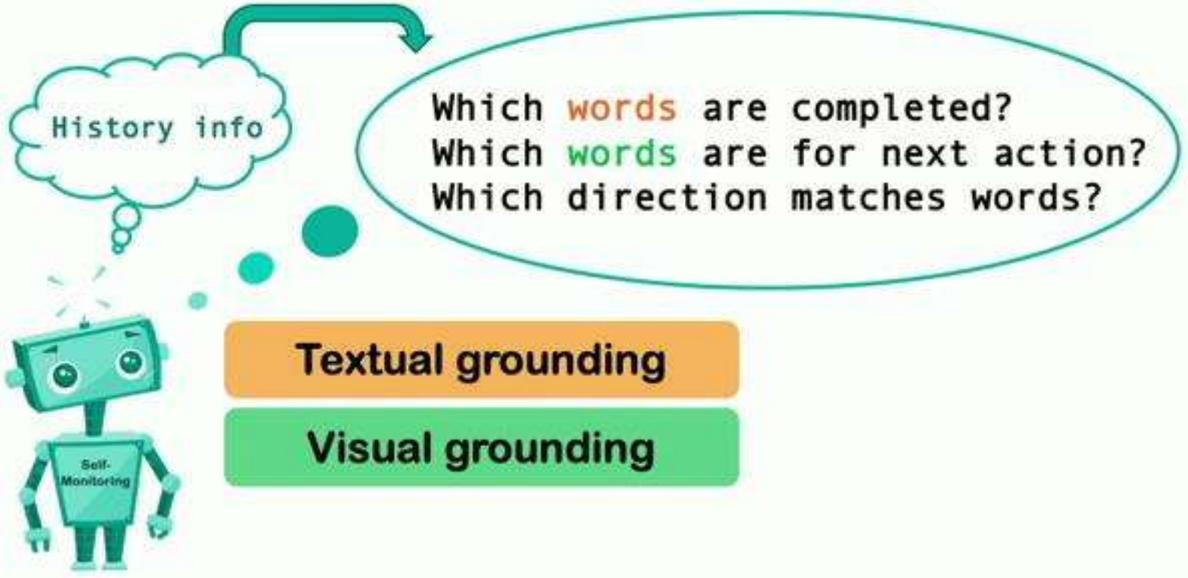
Georgia Tech

Exit the bedroom and go towards the table. Go to the stairs on the left of the couch. Wait on the third step.



Self-Monitoring Navigation Agent

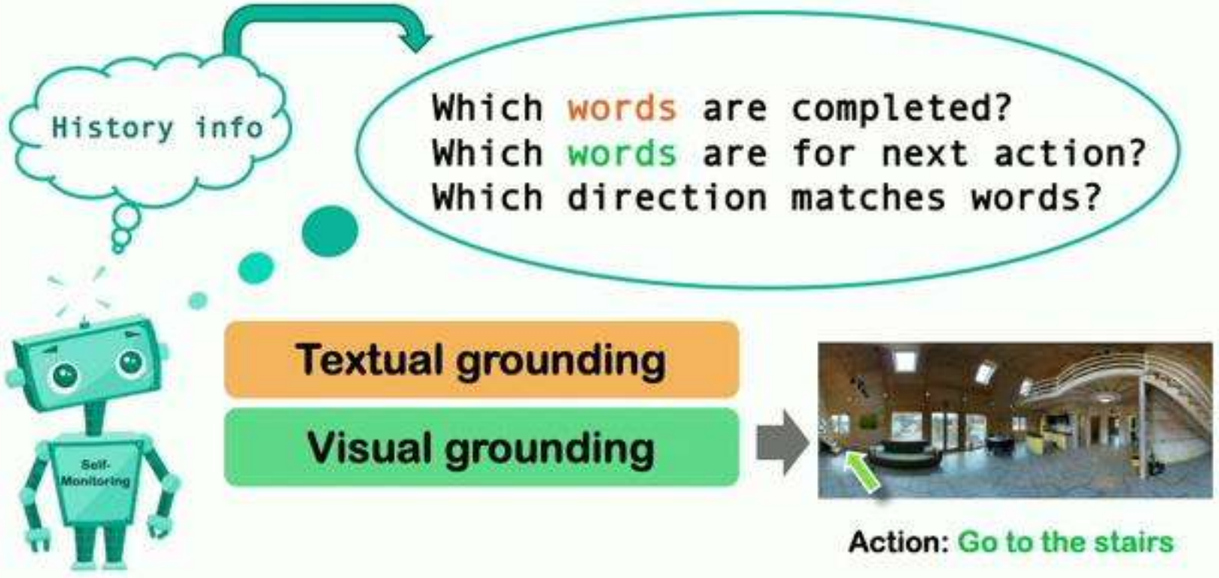
Exit the bedroom and go towards the table. Go to the stairs on the left of the couch. Wait on the third step.



Self-Monitoring Navigation Agent

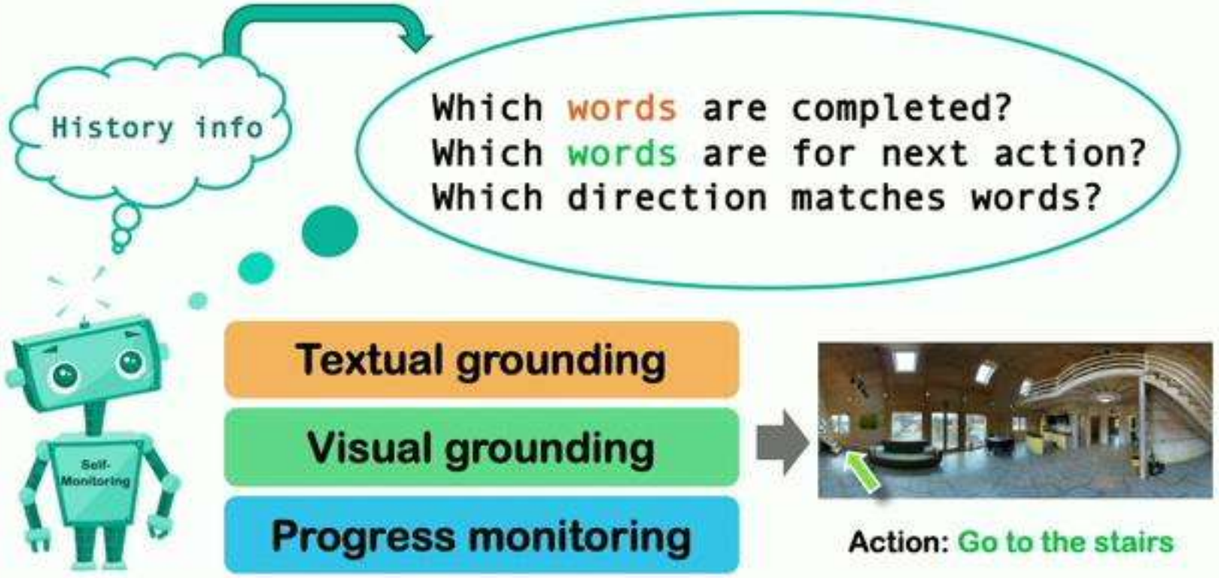
Georgia Tech

Exit the bedroom and go towards the table. Go to the stairs on the left of the couch. Wait on the third step.



Self-Monitoring Navigation Agent

Exit the bedroom and go towards the table. Go to the stairs on the left of the couch. Wait on the third step.



Progress Monitoring

Enforce the grounding on the instruction to be correct.

Self-Monitoring Navigation Agent

Exit the bedroom and go towards the table. Go to the stairs on the left of the couch. Wait on the third step.

The relative position of the grounded instruction implicitly reflects the progress towards the goal.

Self-Monitoring Navigation Agent

Exit the bedroom and go towards the table. Go to the stairs on the left of the couch. Wait on the third step.

The relative position of the grounded instruction implicitly reflects the progress towards the goal.

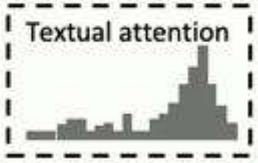
$$p_t^{pm} = f(\alpha_t, \hat{v}_t, h_t, c_t)$$

Attention weights on instruction

Observed image

LSTM hidden states

Textual attention



Self-Monitoring Navigation Agent

Exit the bedroom and go towards the table. Go to the stairs on the left of the couch. Wait on the third step.

The relative position of the grounded instruction implicitly reflects the progress towards the goal.

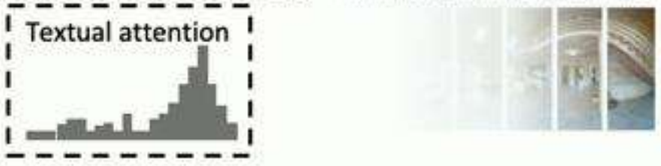
$$p_t^{pm} = f(\alpha_t, \hat{v}_t, h_t, c_t)$$

Attention weights on instruction

Observed image

LSTM hidden states

Textual attention



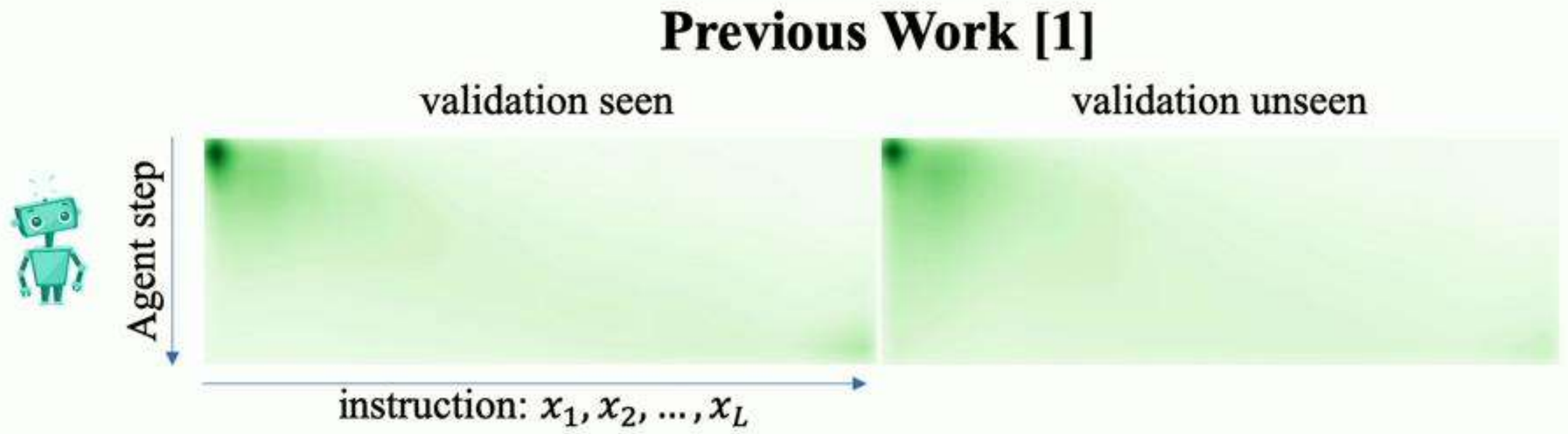
$$\mathcal{L}_{pm} = \sum_{t=1}^{T} (y_t^{pm} - p_t^{pm})^2$$

Progress target

0, at the beginning

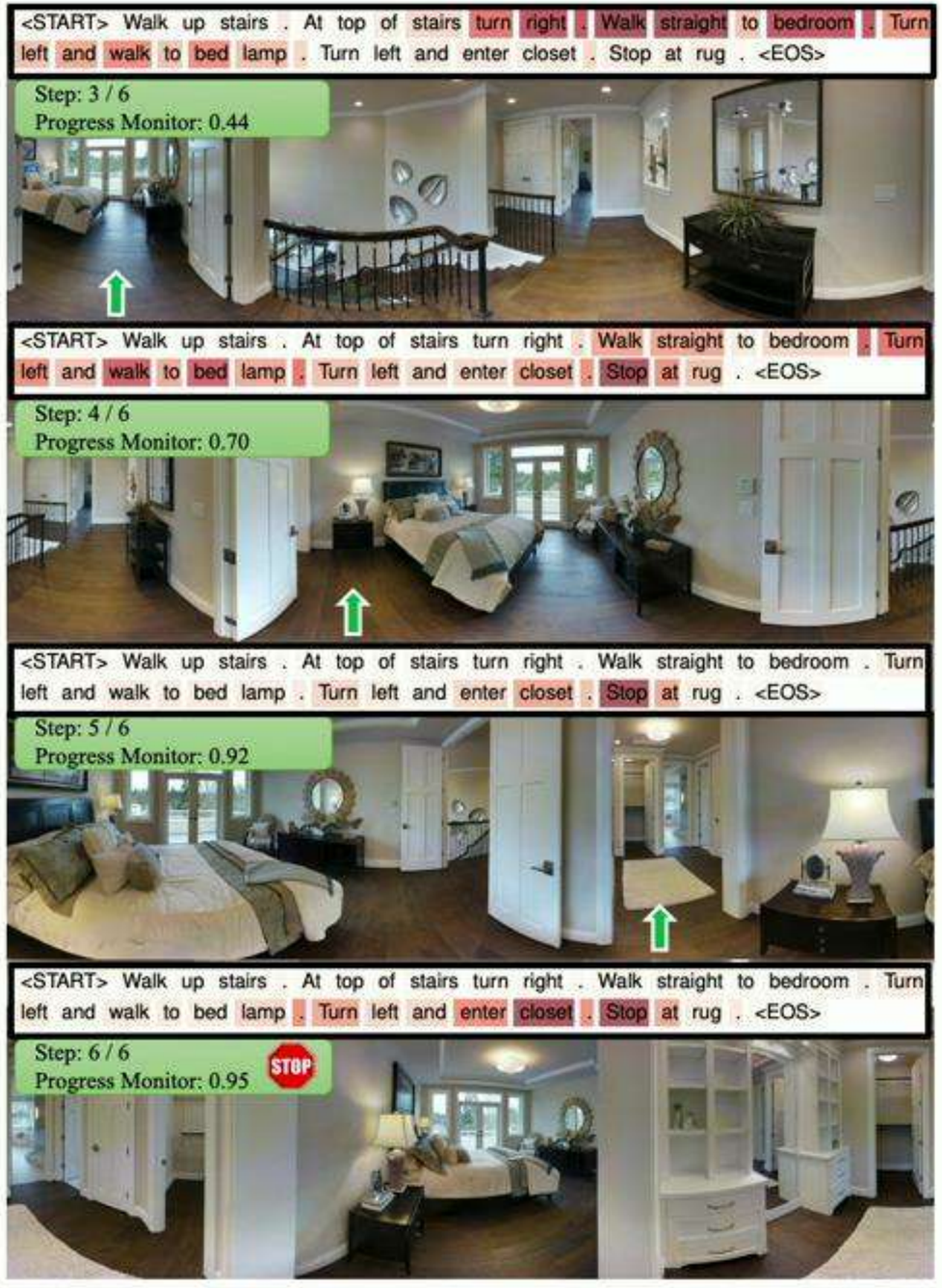
1, reach the goal

Improved Grounding on Instructions

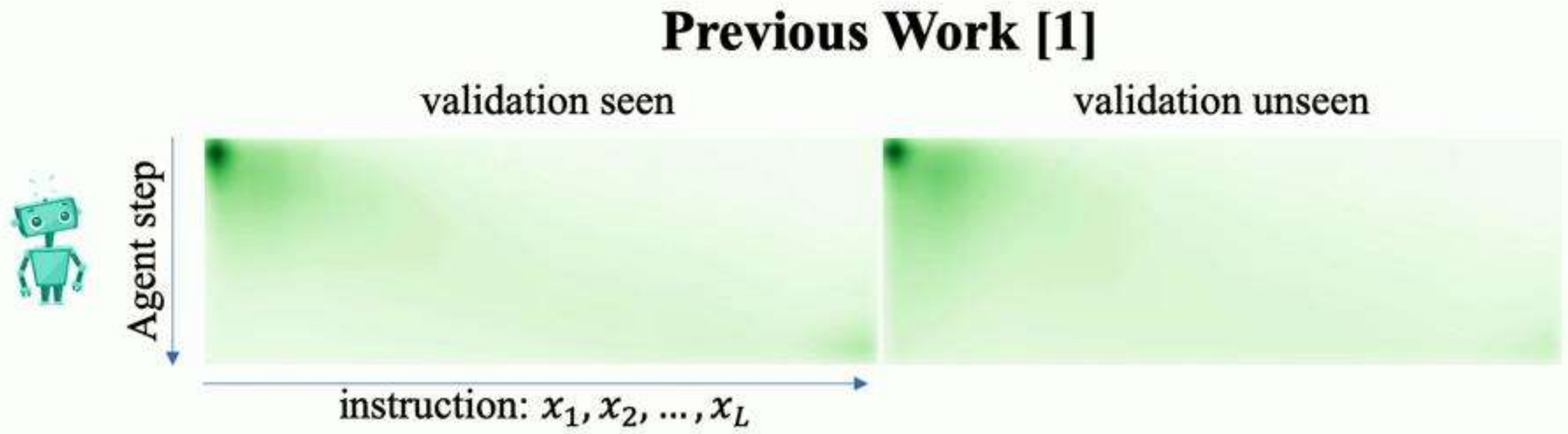


[1] Speaker-Follower (Fried et al., NeurIPS 2018)

Qualitative Results

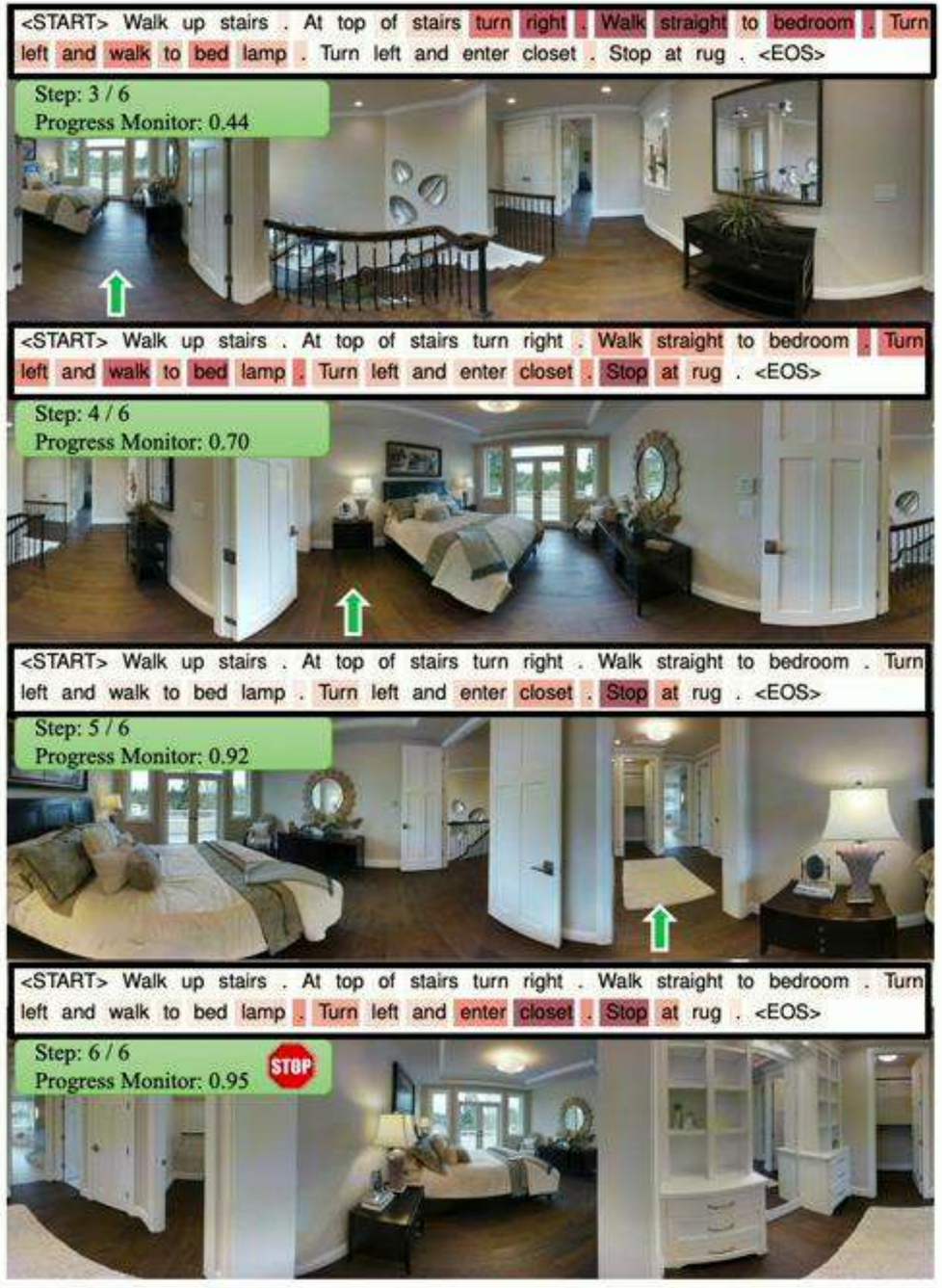


Improved Grounding on Instructions



[1] Speaker-Follower (Fried et al., NeurIPS 2018)

Qualitative Results



Self-Monitoring Navigation Agent

Exit the bedroom and go towards the table. Go to the stairs on the left of the couch. Wait on the third step.

The relative position of the grounded instruction implicitly reflects the progress towards the goal.

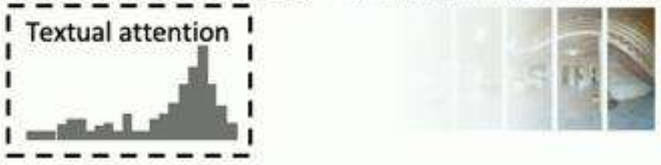
$$p_t^{pm} = f(\alpha_t, \hat{v}_t, h_t, c_t)$$

Attention weights on instruction

Observed image

LSTM hidden states

Textual attention



$$\mathcal{L}_{pm} = \sum_{t=1}^{T} (y_t^{pm} - p_t^{pm})^2$$

Progress target

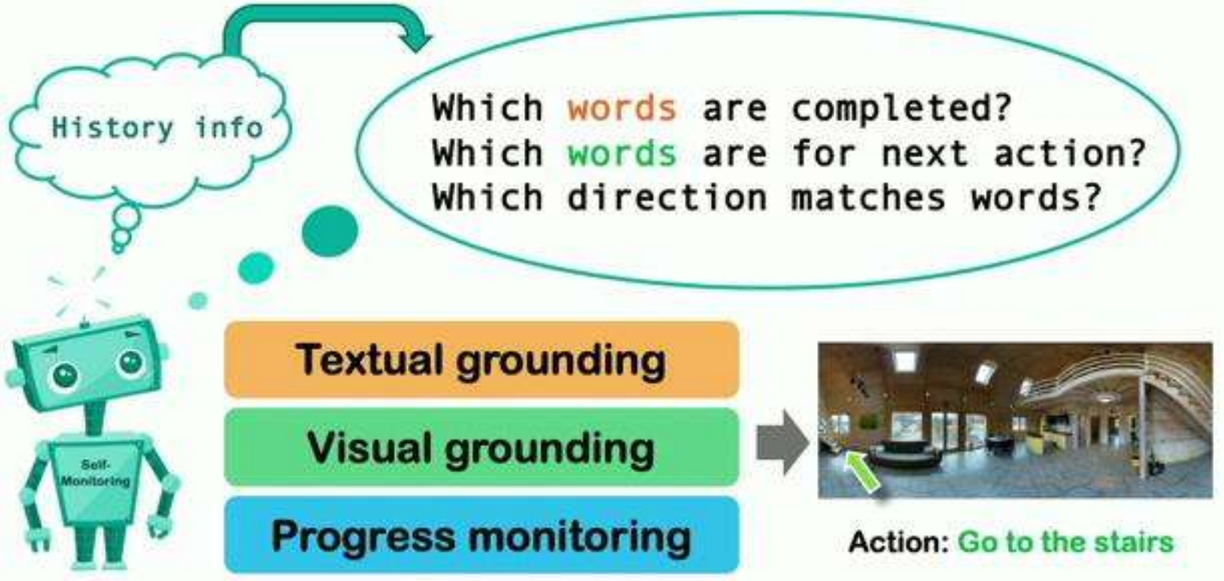
0, at the beginning

1, reach the goal

Self-Monitoring Navigation Agent

Georgia Tech

Exit the bedroom and go towards the table. Go to the stairs on the left of the couch. Wait on the third step.



Progress Monitoring

Enforce the grounding on the instruction to be correct.

Self-Monitoring Navigation Agent

Exit the bedroom and go towards the table. Go to the stairs on the left of the couch. Wait on the third step.

The relative position of the grounded instruction implicitly reflects the progress towards the goal.

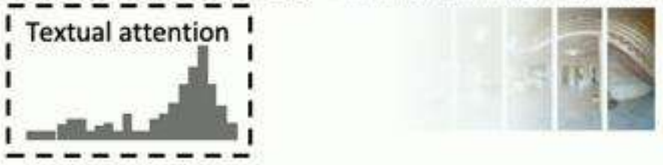
$$p_t^{pm} = f(\alpha_t, \hat{v}_t, h_t, c_t)$$

Attention weights on instruction

Observed image

LSTM hidden states

Textual attention



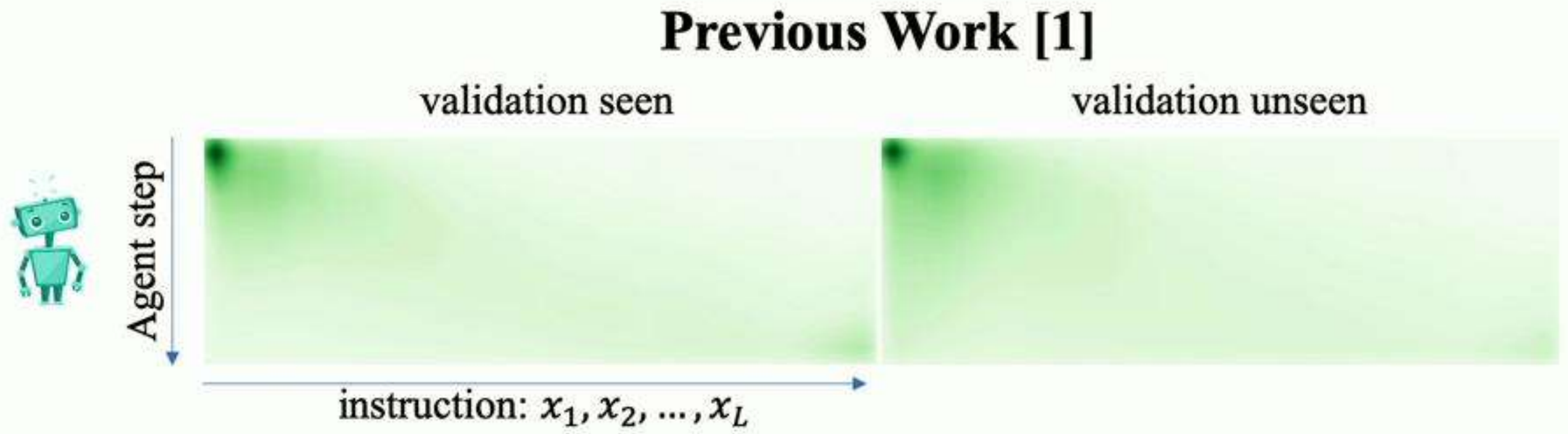
$$\mathcal{L}_{pm} = \sum_{t=1}^{T} (y_t^{pm} - p_t^{pm})^2$$

Progress target

0, at the beginning

1, reach the goal

Improved Grounding on Instructions



[1] Speaker-Follower (Fried et al., NeurIPS 2018)

Self-Monitoring Navigation Agent

Exit the bedroom and go towards the table. Go to the stairs on the left of the couch. Wait on the third step.

The relative position of the grounded instruction implicitly reflects the progress towards the goal.

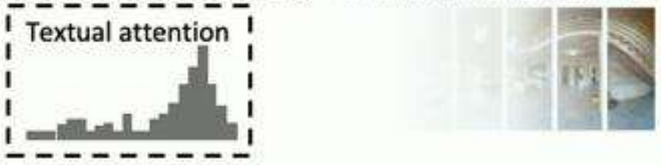
$$p_t^{pm} = f(\alpha_t, \hat{v}_t, h_t, c_t)$$

Attention weights on instruction

Observed image

LSTM hidden states

Textual attention



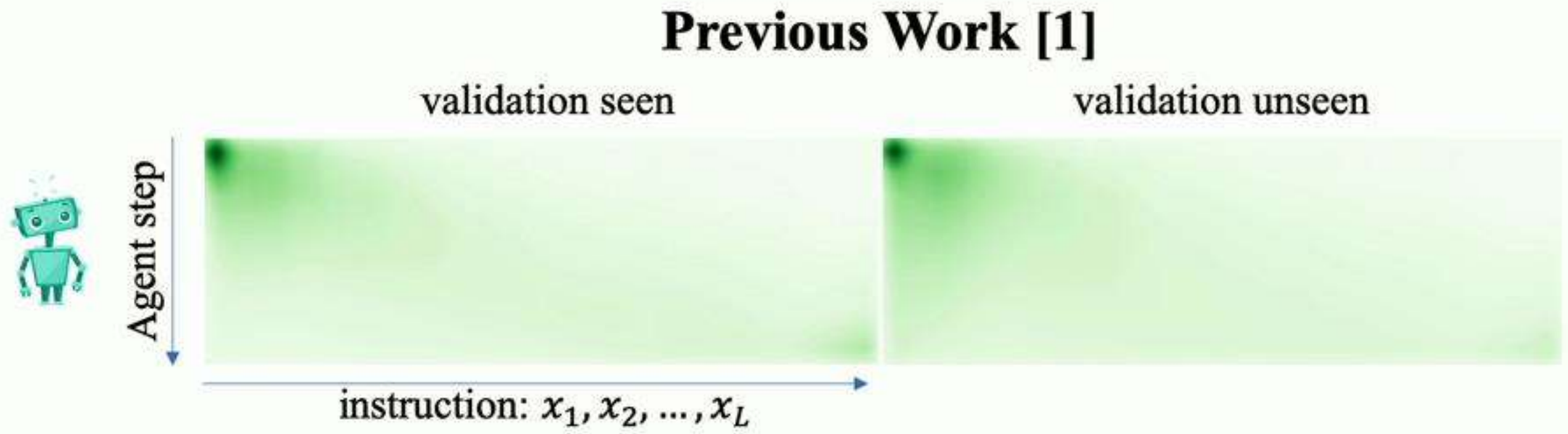
$$\mathcal{L}_{pm} = \sum_{t=1}^{T} (y_t^{pm} - p_t^{pm})^2$$

Progress target

0, at the beginning

1, reach the goal

Improved Grounding on Instructions



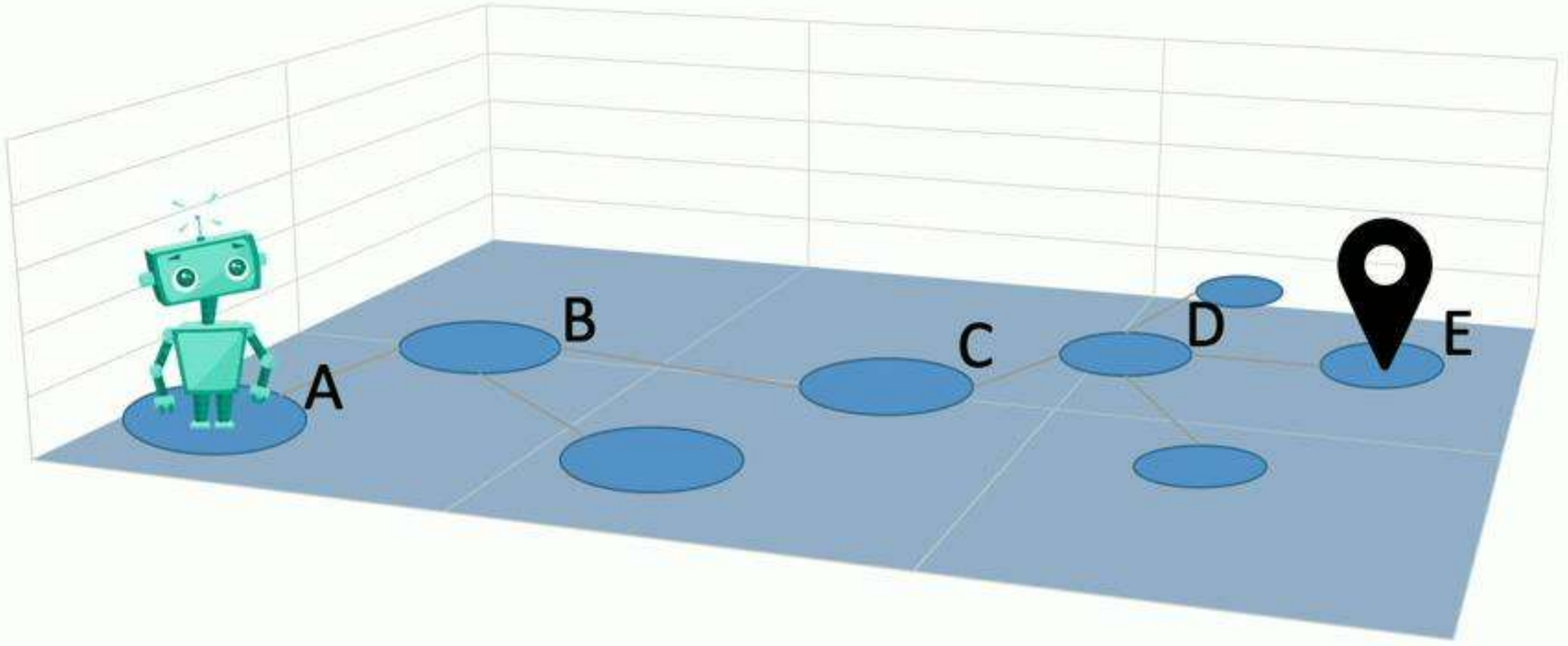
[1] Speaker-Follower (Fried et al., NeurIPS 2018)

Qualitative Results

Self-Monitoring Agent - Progress Estimation

Georgia Tech

Move to B. Go to C. Exist to D. Stop at E.

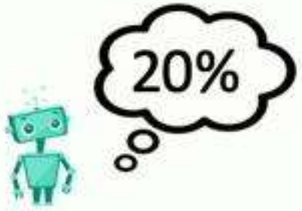


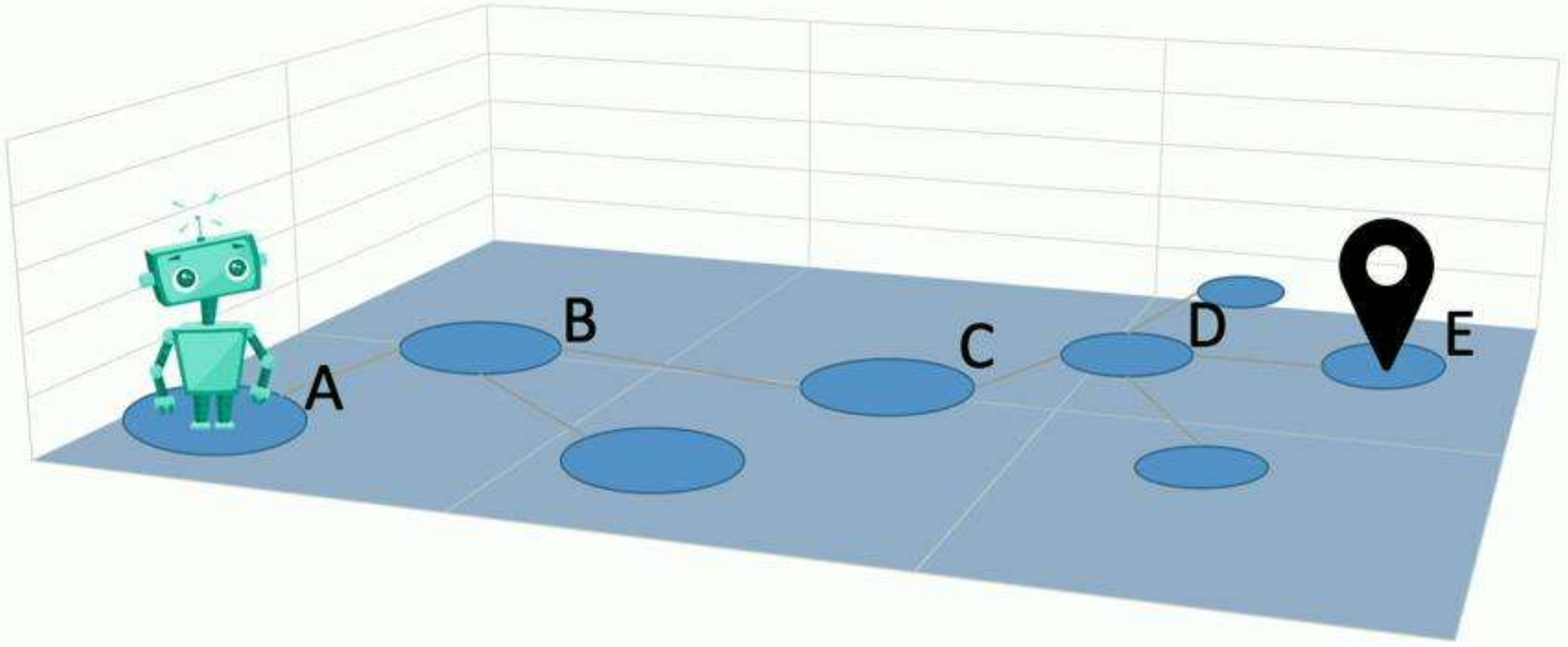
Self-Monitoring Agent - Progress Estimation

Georgia Tech

Textual grounding

Move to B. Go to C. Exist to D. Stop at E.





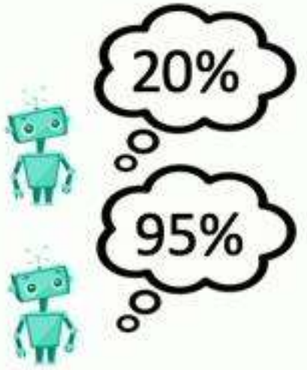
Self-Monitoring Agent - Progress Estimation

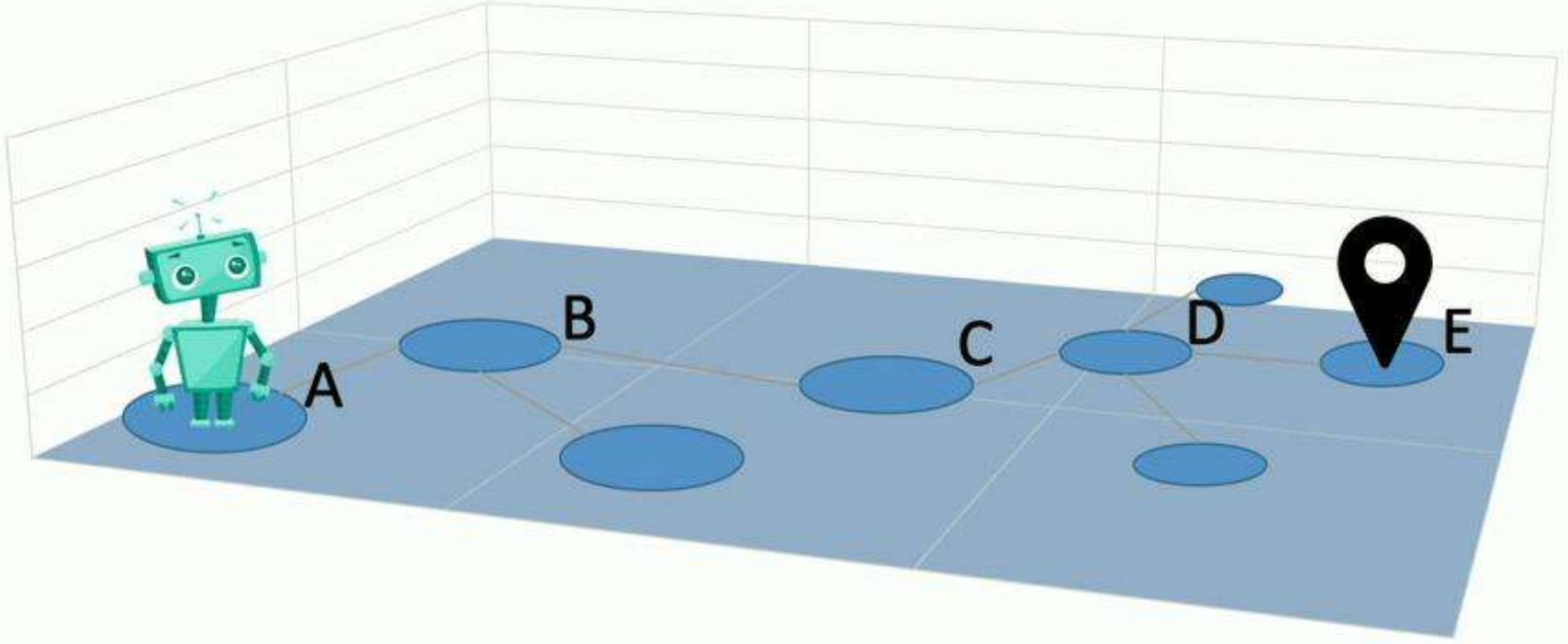
Georgia Tech

Textual grounding

Move to B. Go to C. Exist to D. Stop at E.

Move to B. Go to C. Exist to D. Stop at E.





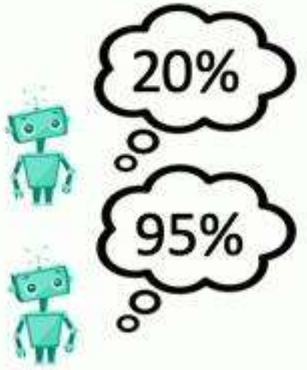
Self-Monitoring Agent - Progress Estimation

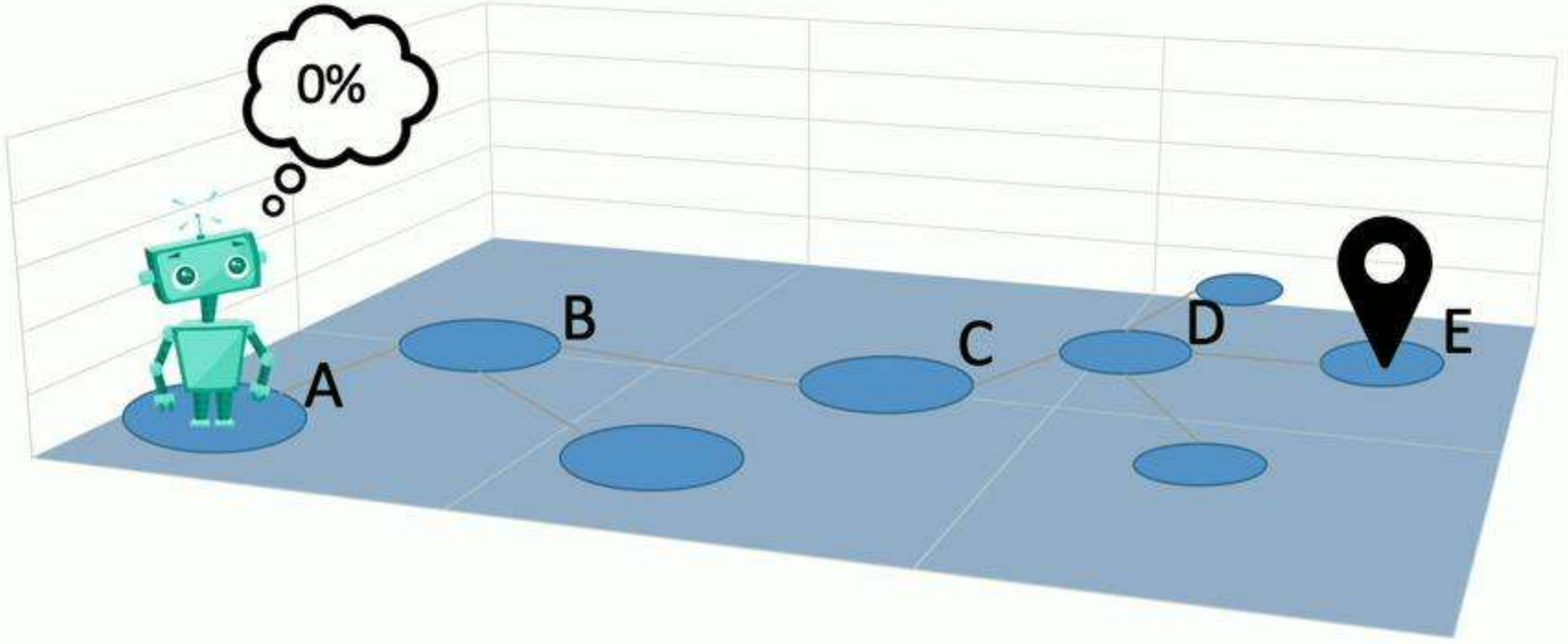
Georgia Tech

Textual grounding

Move to B. Go to C. Exist to D. Stop at E.

Move to B. Go to C. Exist to D. Stop at E.





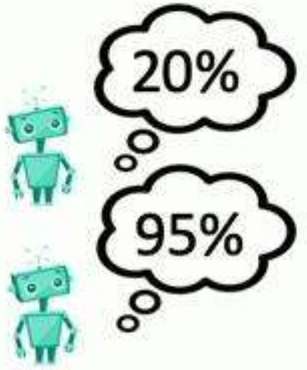
Self-Monitoring Agent - Progress Estimation

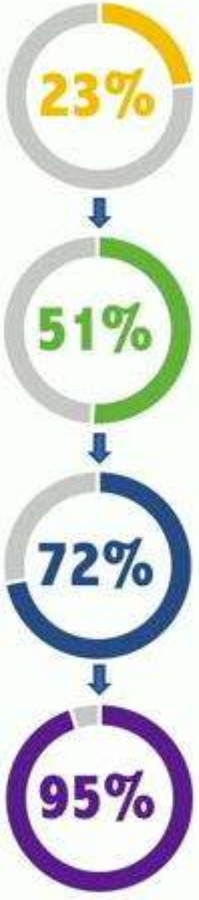
Georgia Tech

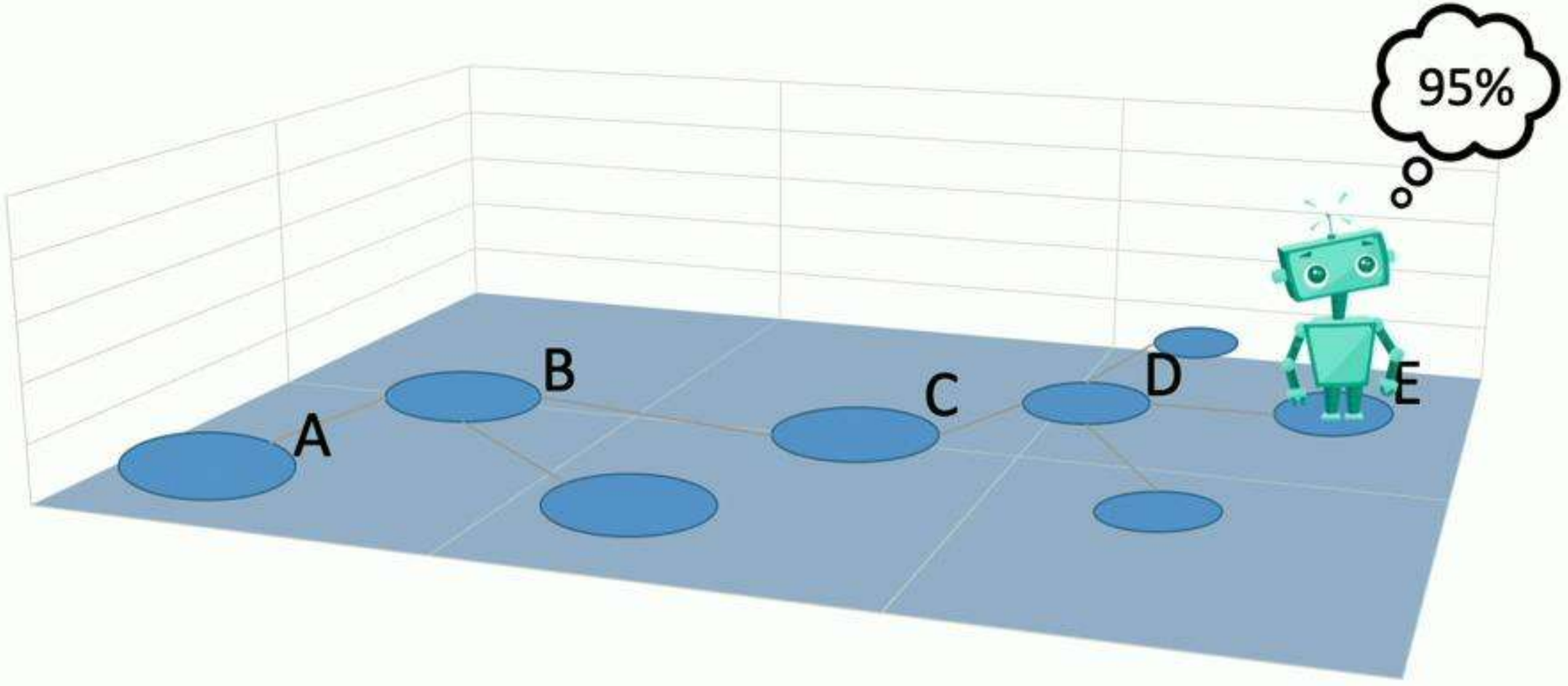
Textual grounding

Move to B. Go to C. Exist to D. Stop at E.

Move to B. Go to C. Exist to D. Stop at E.

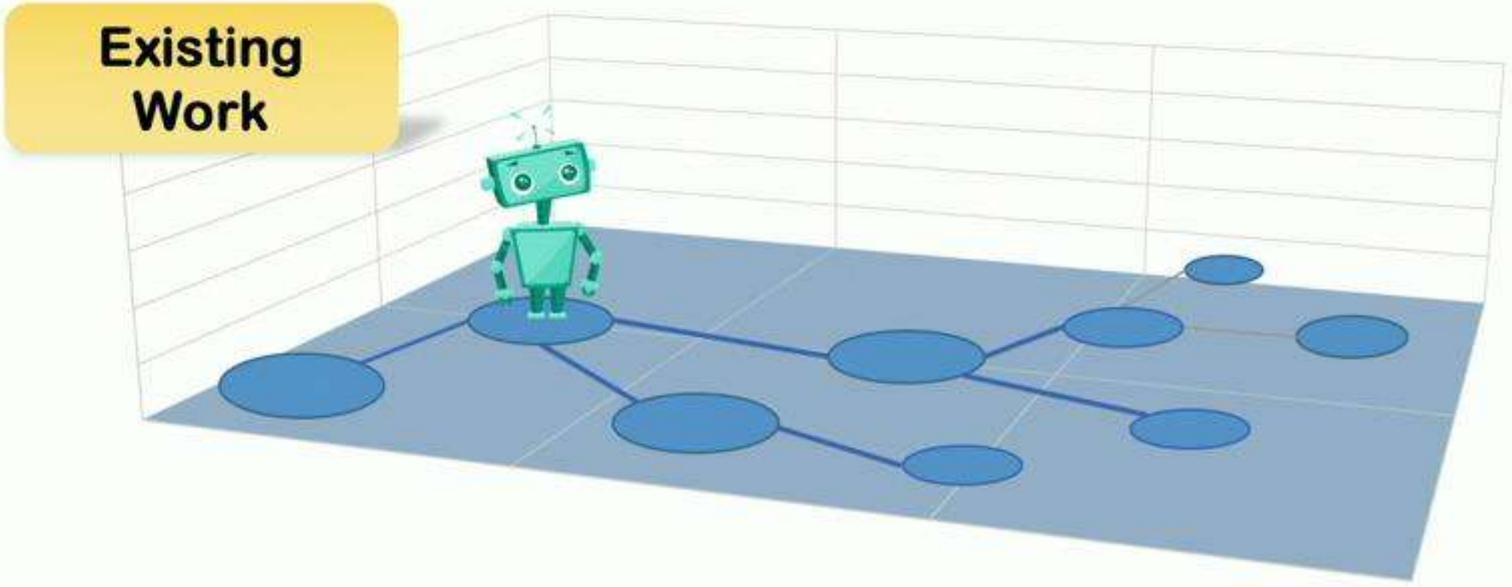






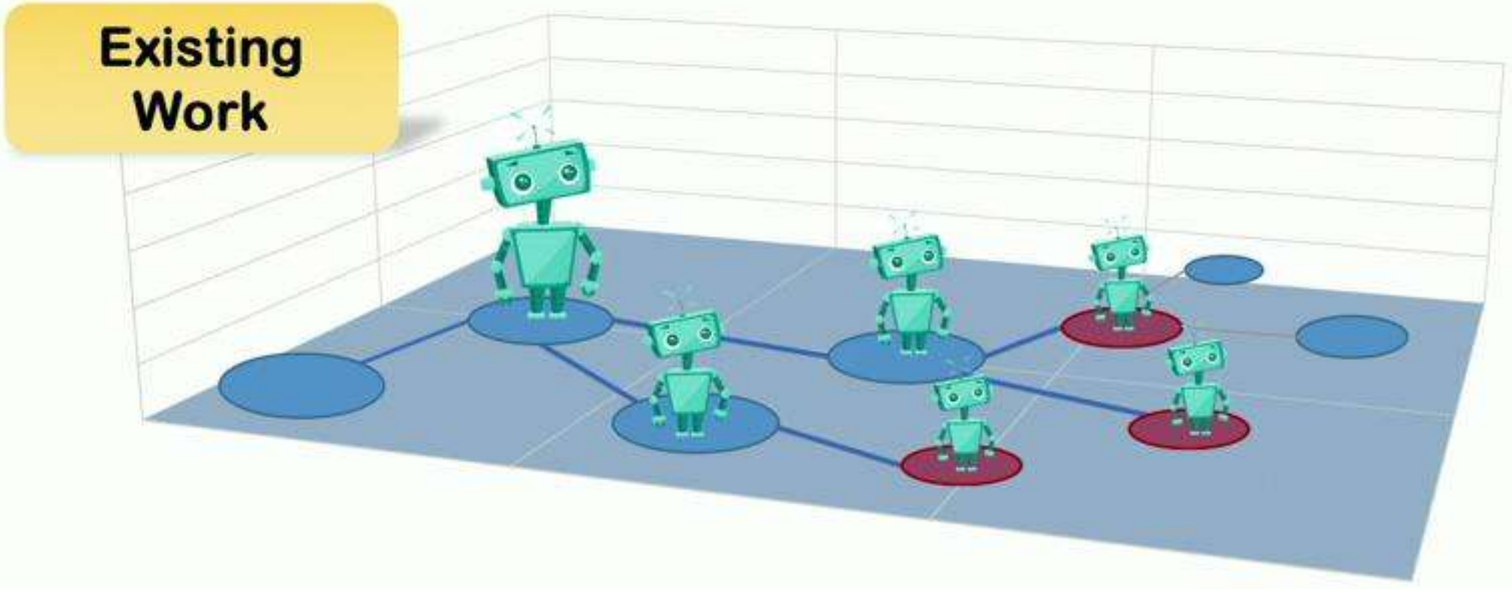
Motivation - The Regretful Agent

- View as navigation graph
  - Existing work uses beam search or maintain frontiers



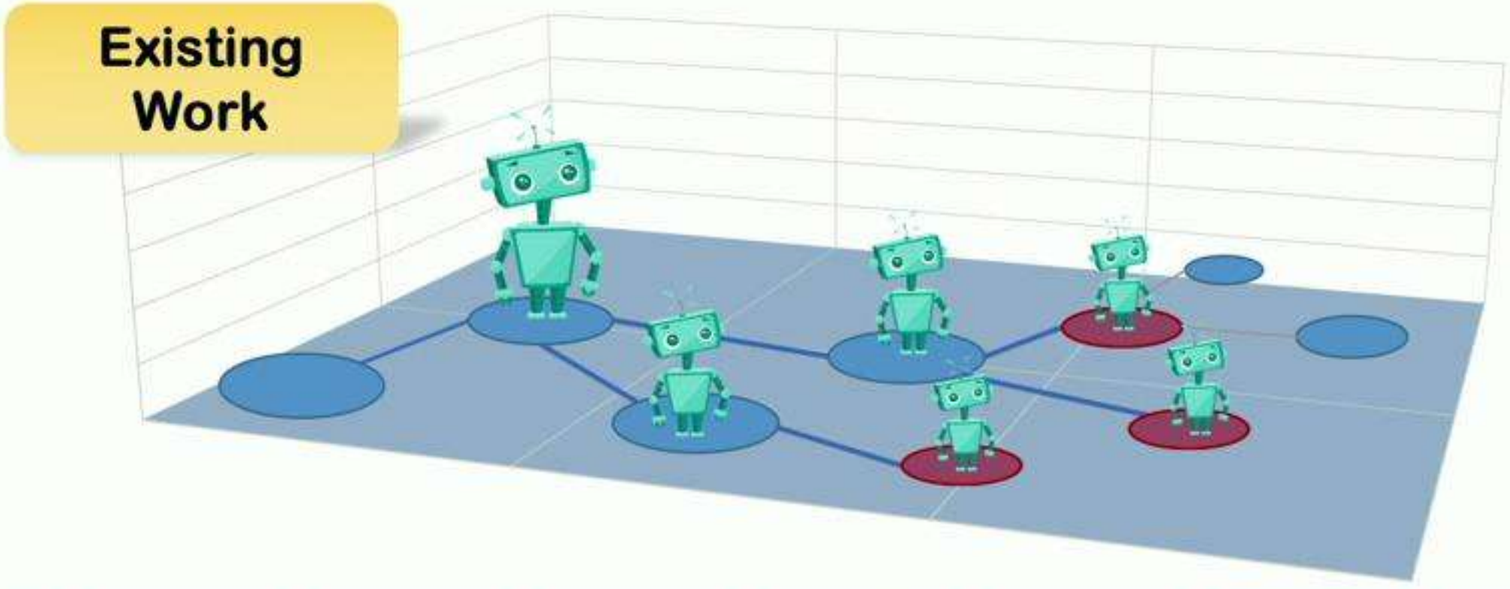
Motivation - The Regretful Agent

- View as navigation graph
  - Existing work uses beam search or maintain frontiers



Motivation - The Regretful Agent

- View as navigation graph
  - Existing work uses beam search or maintain frontiers

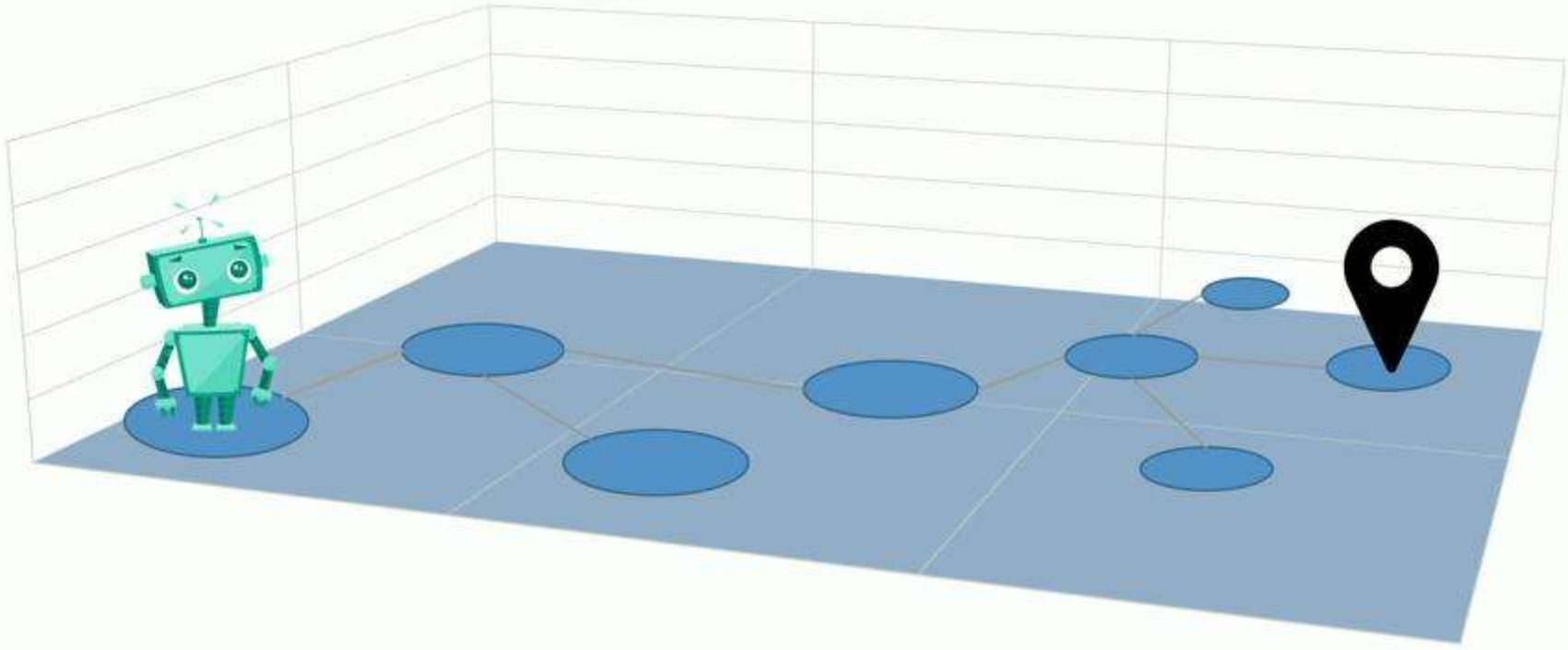


Progress Estimation as learned heuristic function

- provides an informed way to guess which direction is more likely to lead to the goal.

The Regretful Agent

- End-to-End Learned Backtracking Agent (purely greedy)

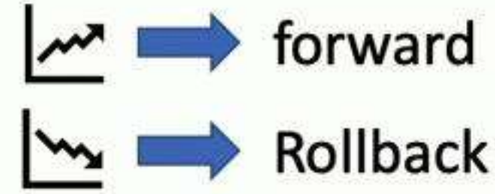


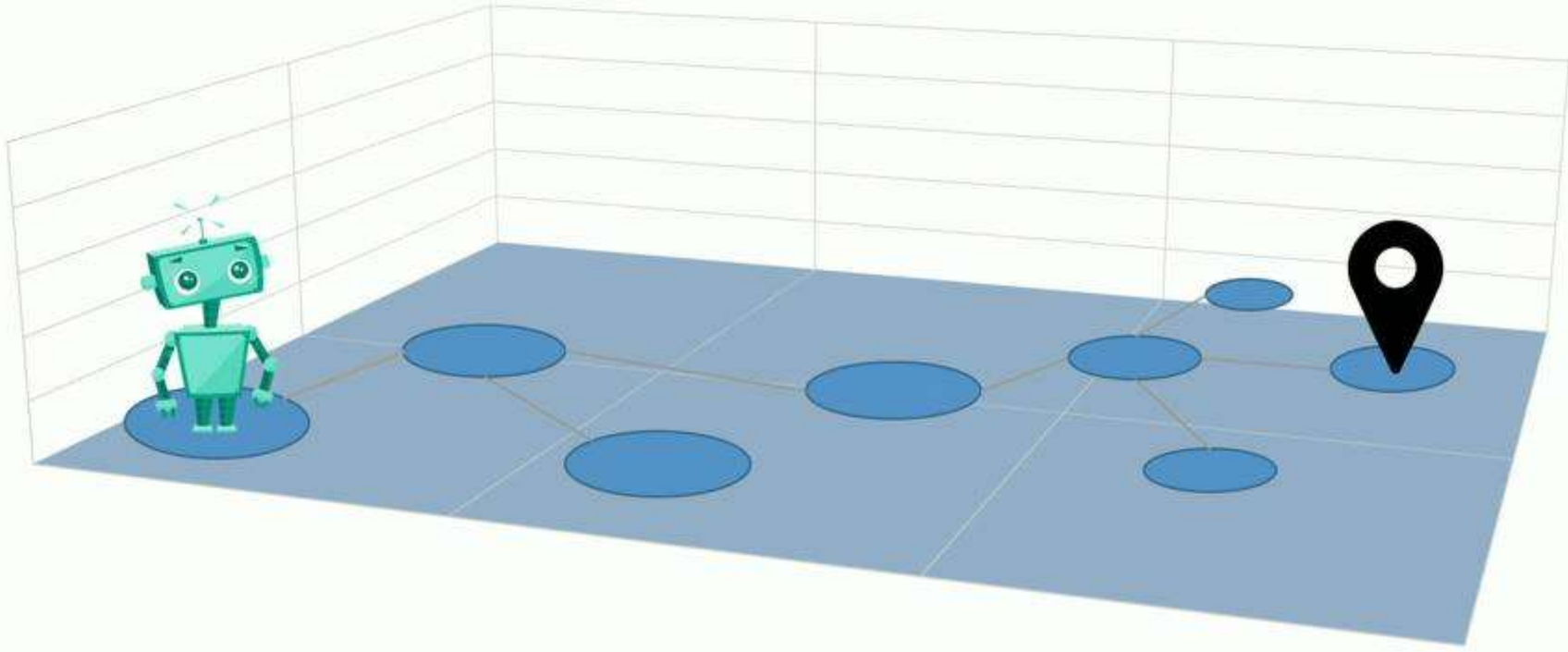
The Regretful Agent

- End-to-End Learned Backtracking Agent (purely greedy)

Regret Module

forward or rollback?





The Regretful Agent

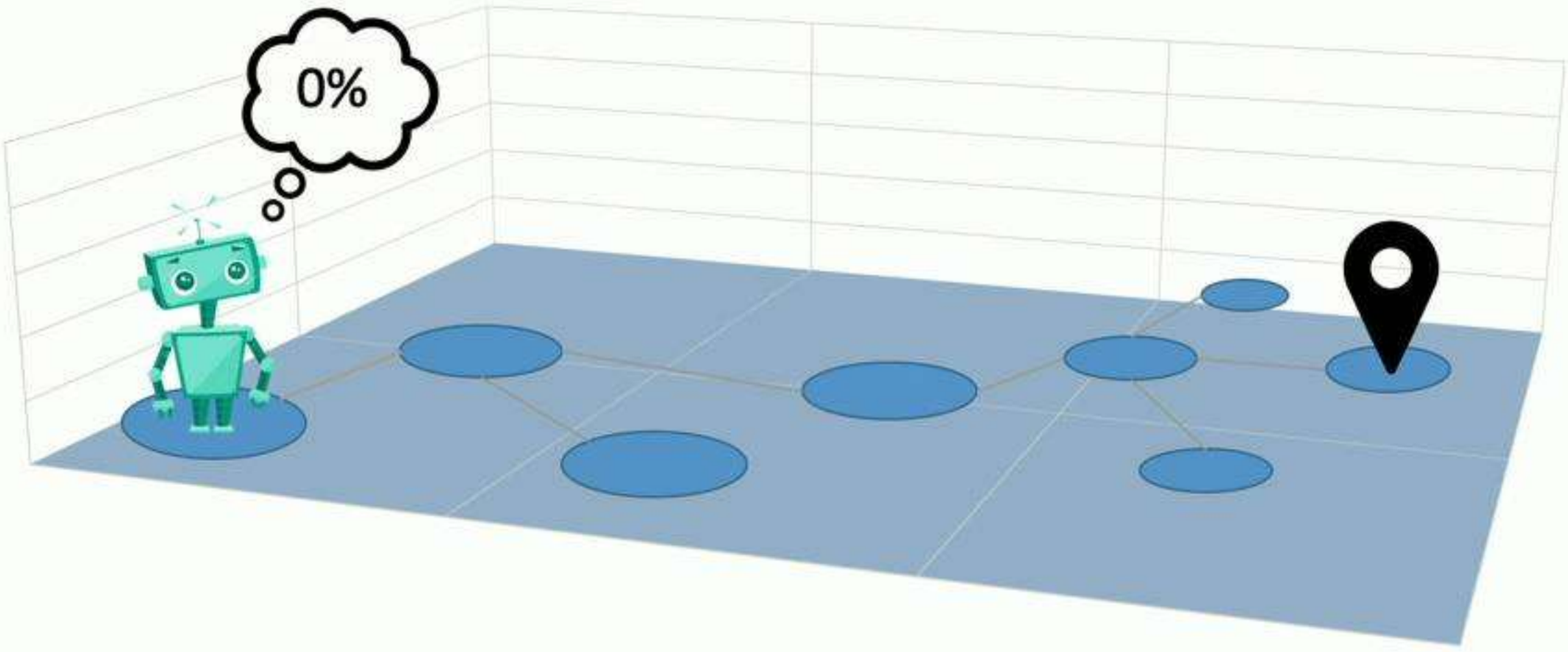
- End-to-End Learned Backtracking Agent (purely greedy)

Regret Module

forward or rollback?

forward

Rollback



The Regretful Agent

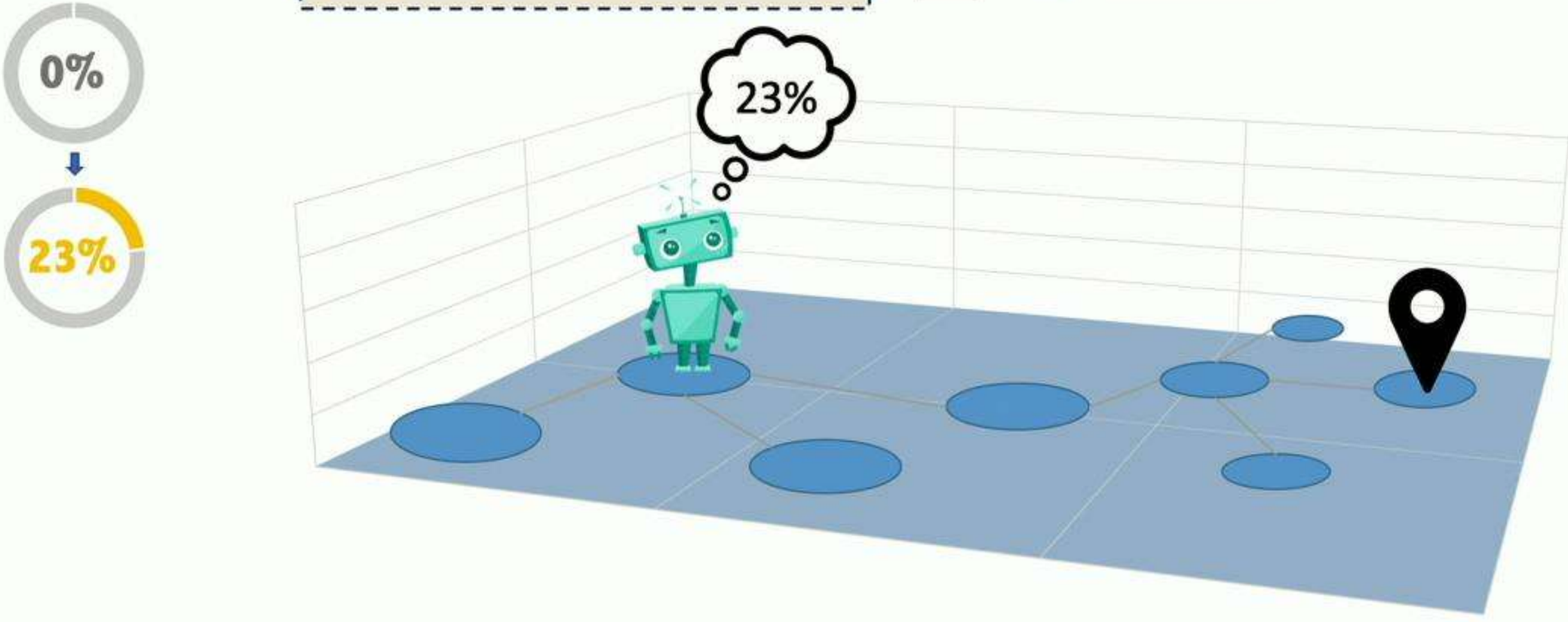
- End-to-End Learned Backtracking Agent (purely greedy)

Regret Module

forward or rollback?

forward

Rollback



The Regretful Agent

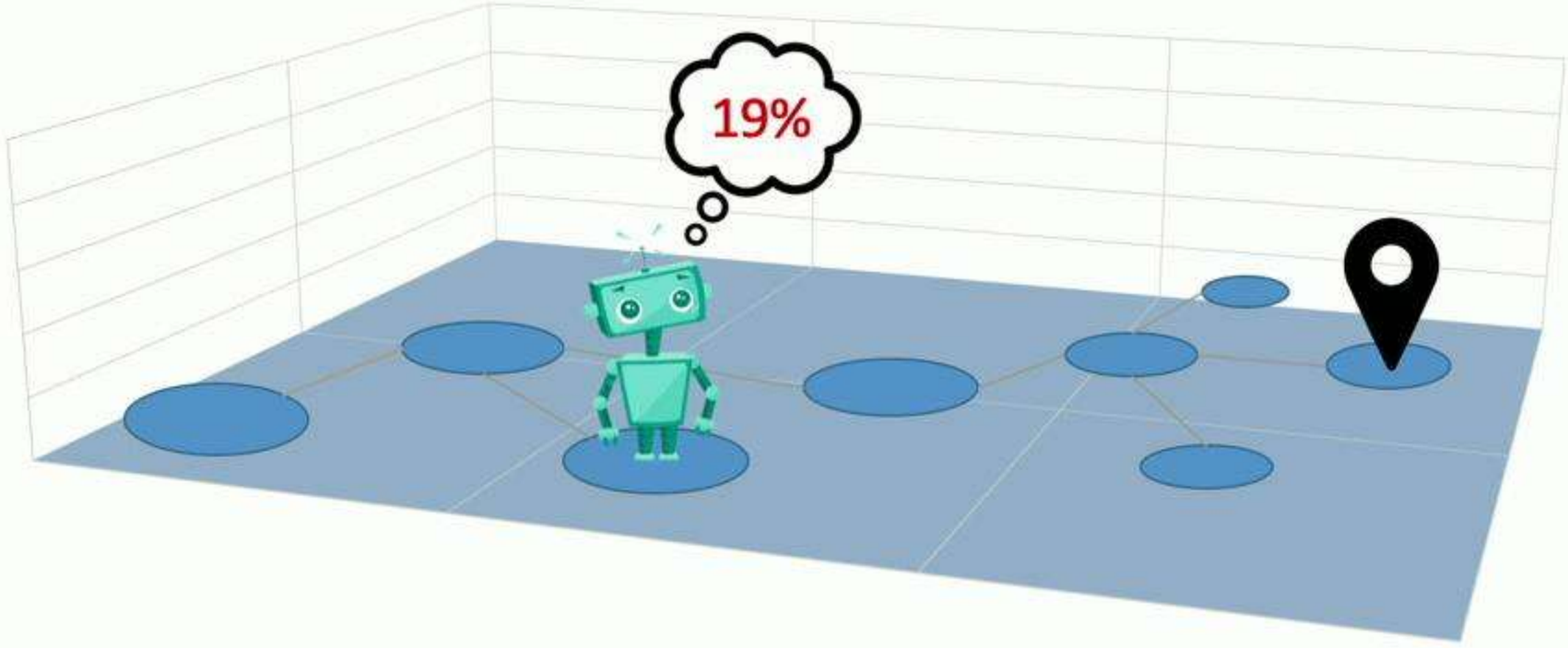
- End-to-End Learned Backtracking Agent (purely greedy)

Regret Module

forward or rollback?

forward

Rollback

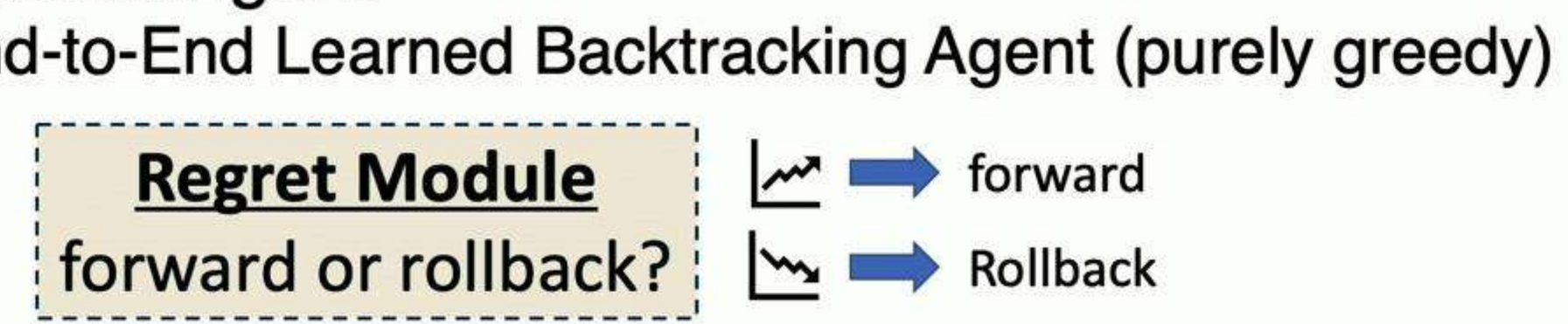


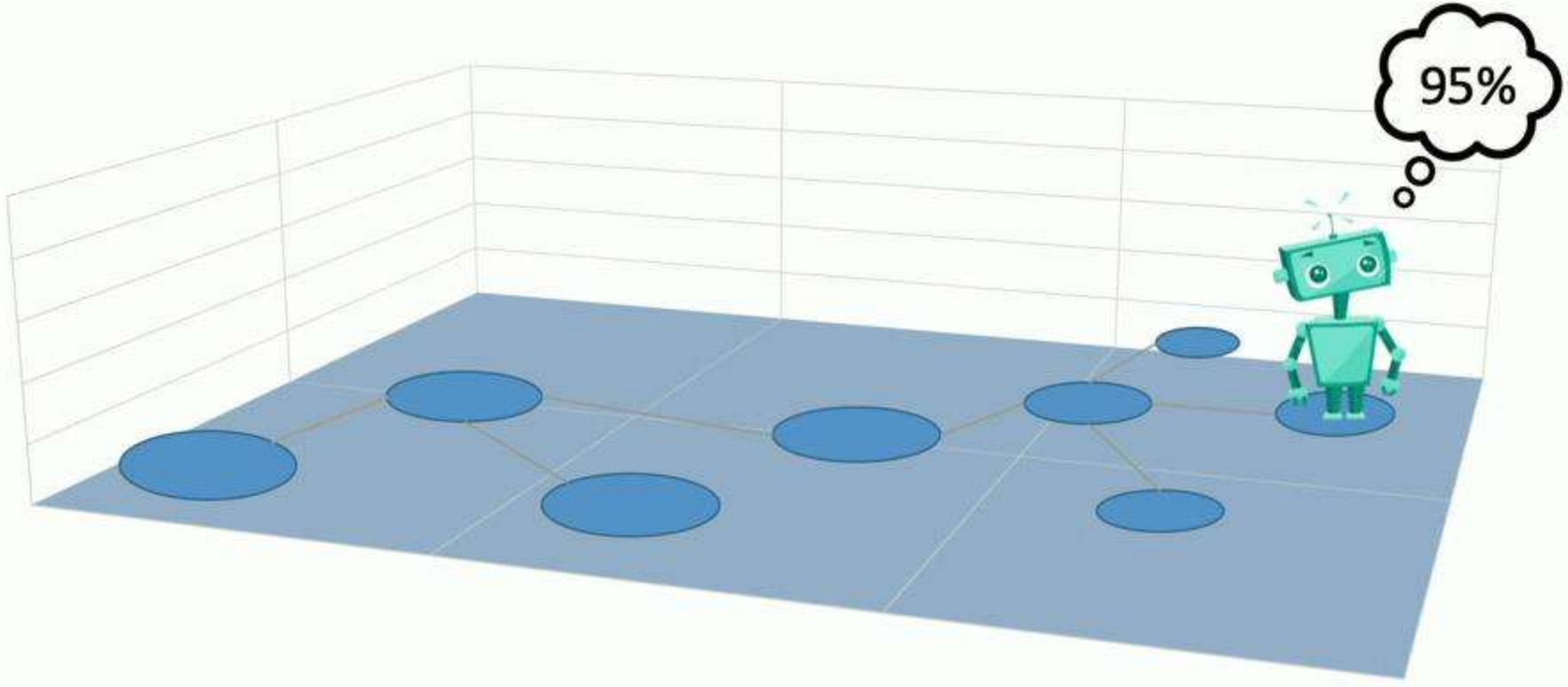
The Regretful Agent

- End-to-End Learned Backtracking Agent (purely greedy)

Regret Module

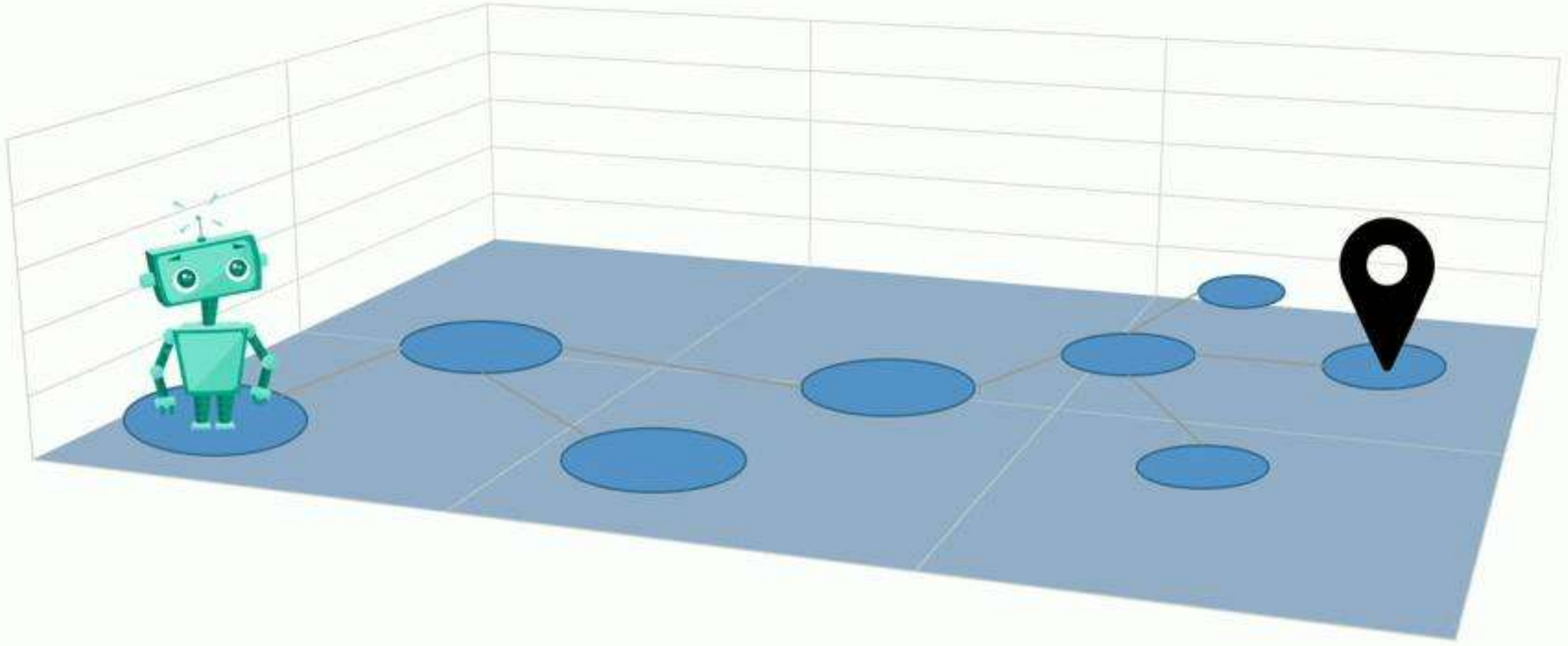
forward or rollback?





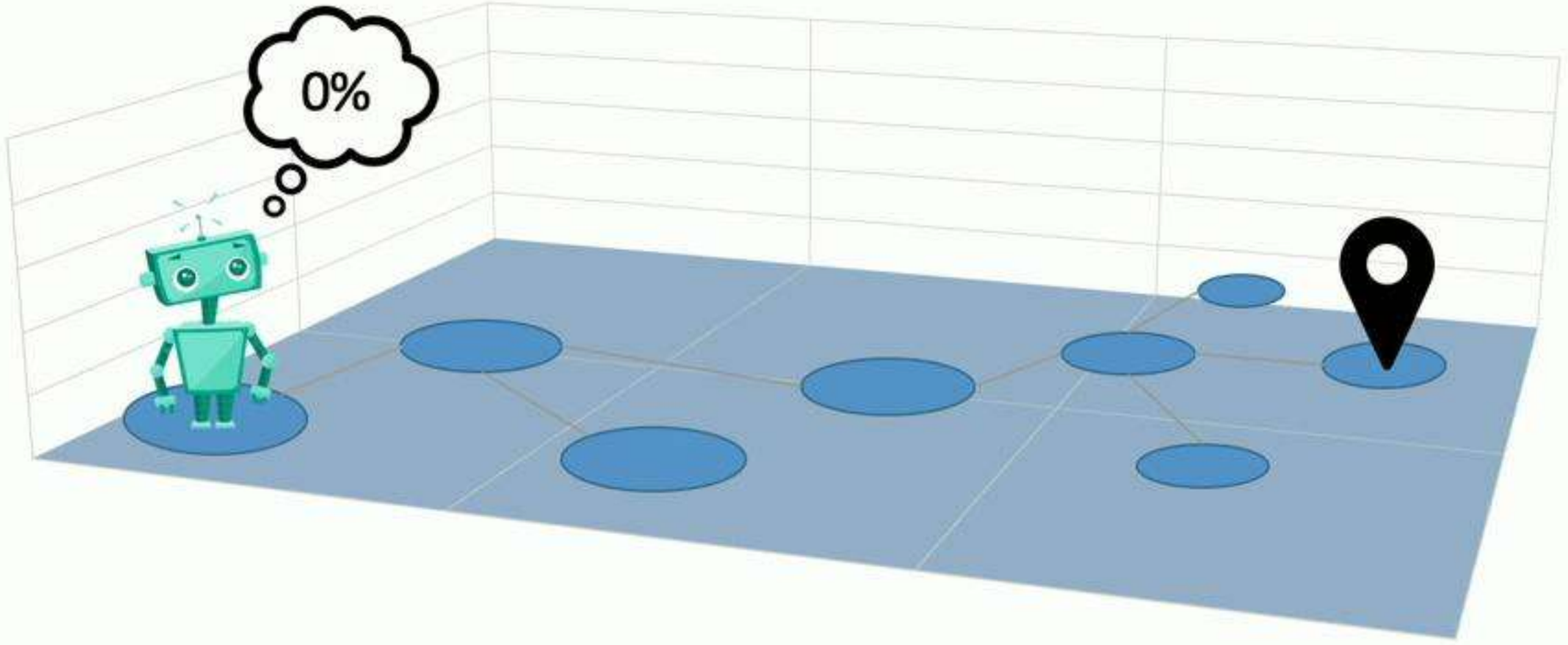
The Regretful Agent – Progress Marker

Progress Marker: which way to go?



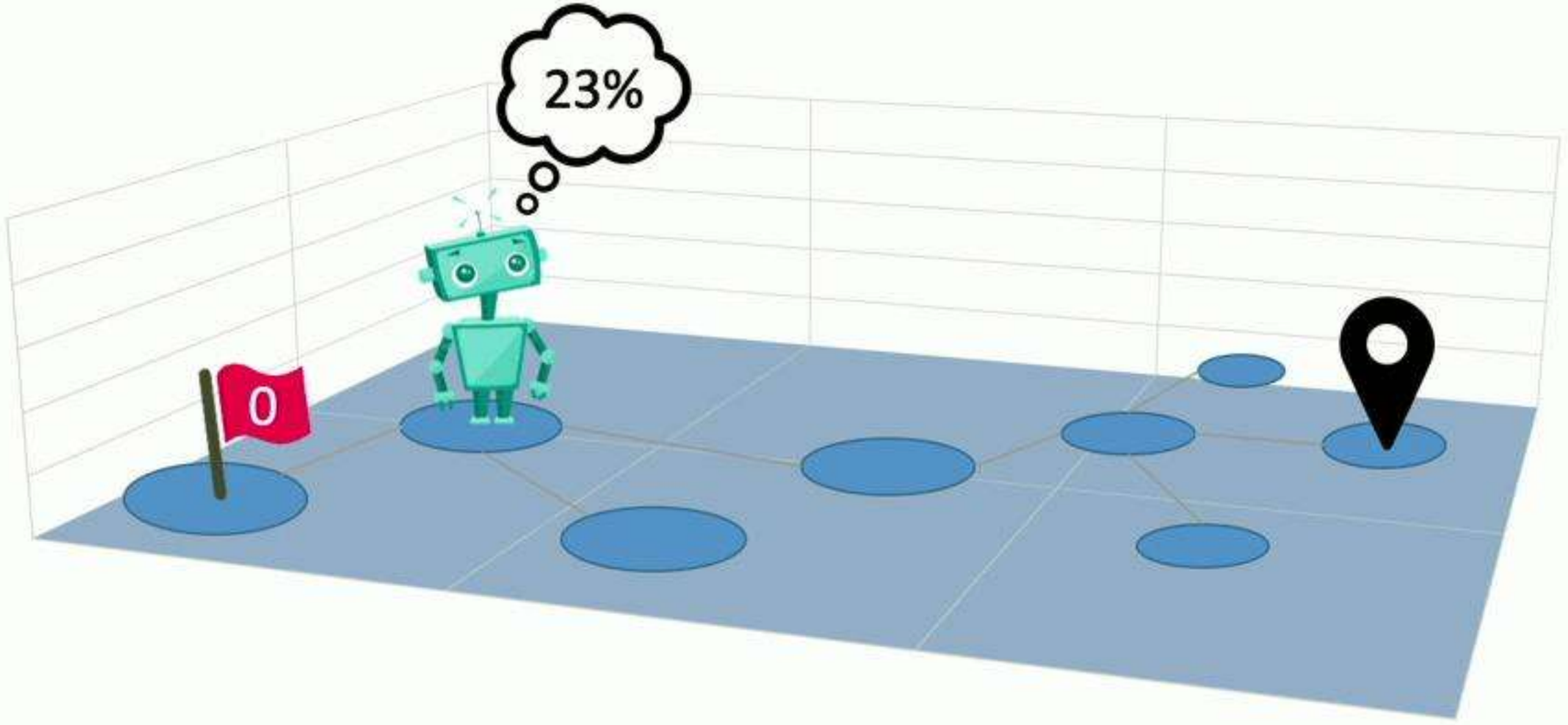
The Regretful Agent – Progress Marker

Progress Marker: which way to go?



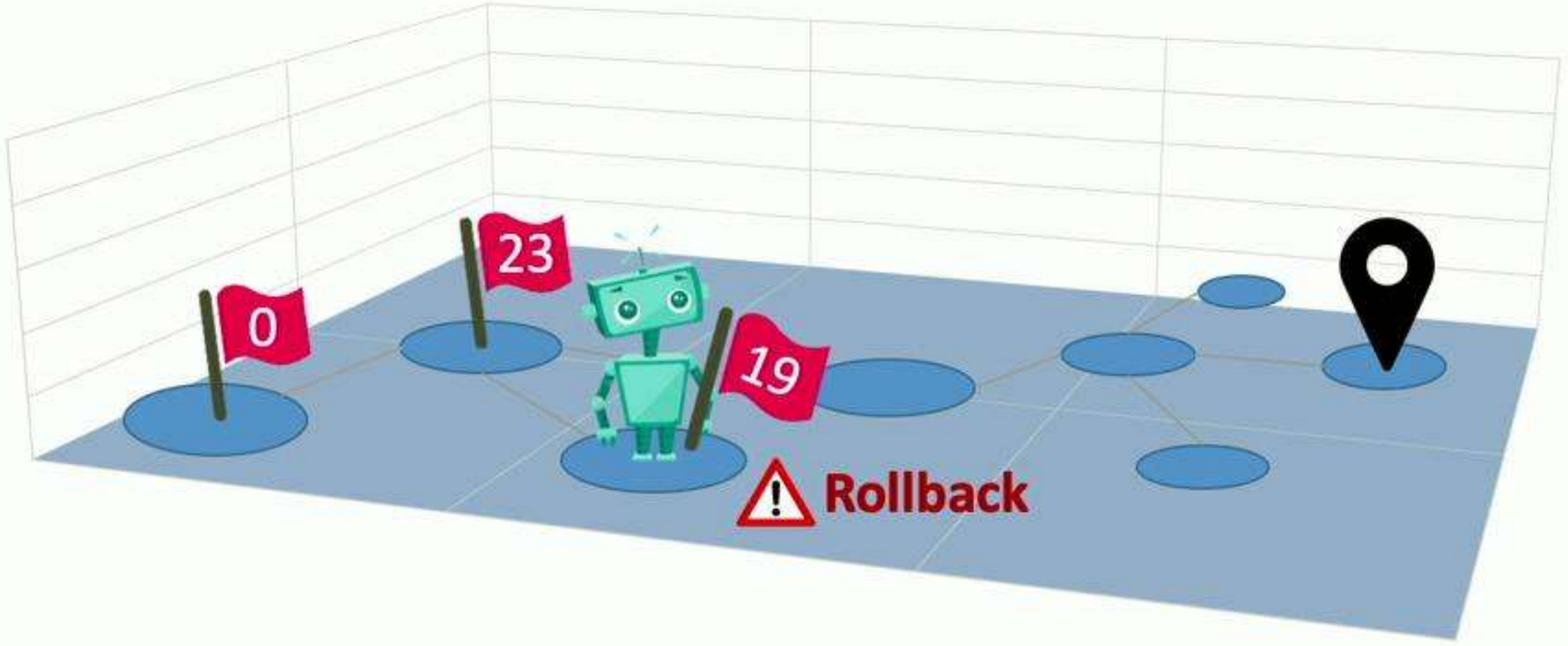
The Regretful Agent – Progress Marker

Progress Marker: which way to go?



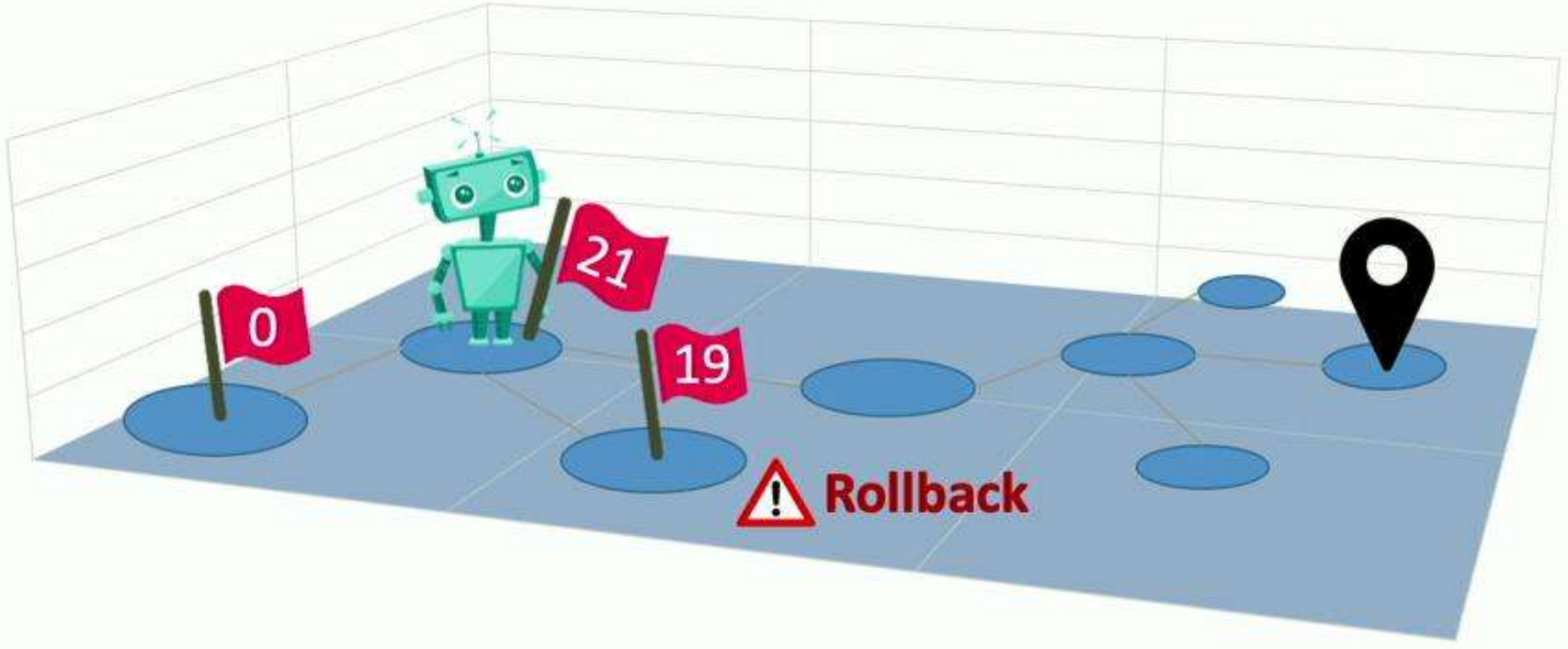
The Regretful Agent – Progress Marker

Progress Marker: which way to go?



The Regretful Agent – Progress Marker

Progress Marker: which way to go?

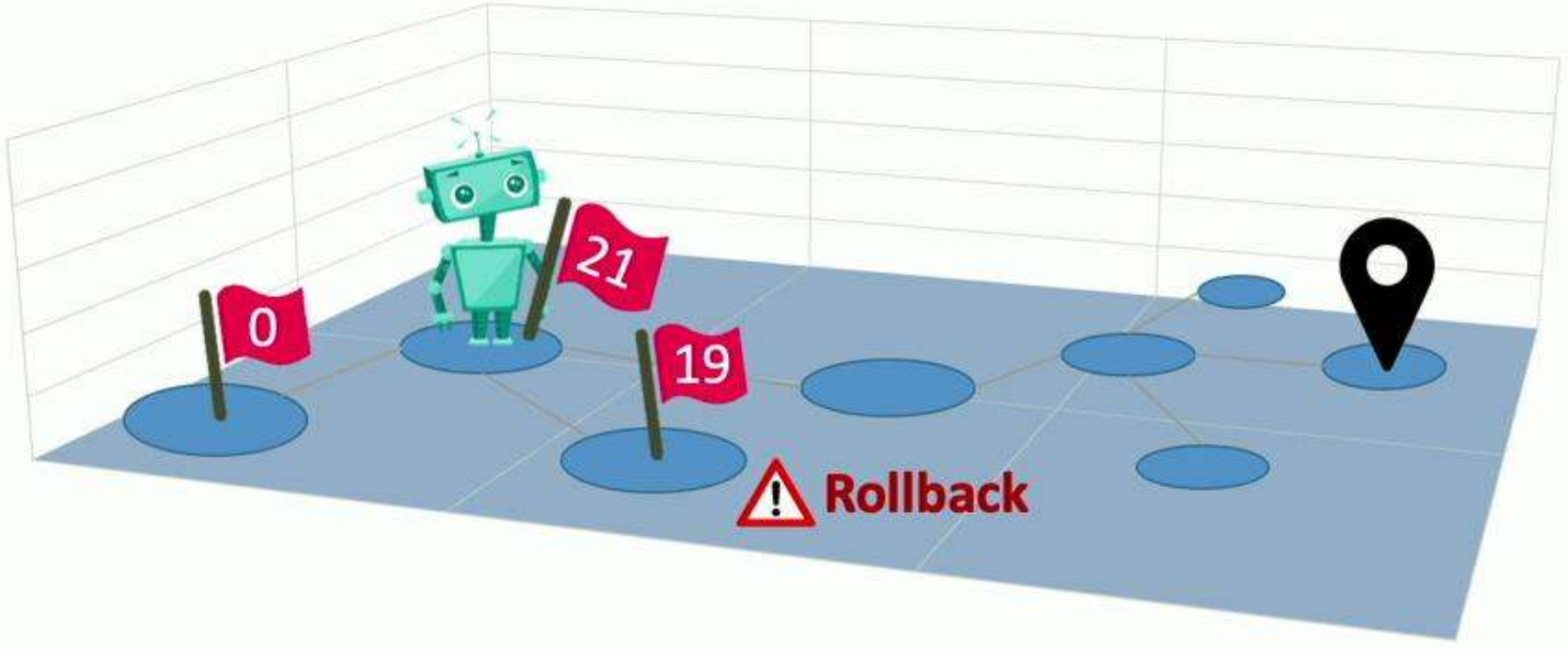


The Regretful Agent – Progress Marker

Progress Marker: which way to go?

Local Graph Search:

- know which directions have been visited
- estimate which one is likely to lead to the goal.

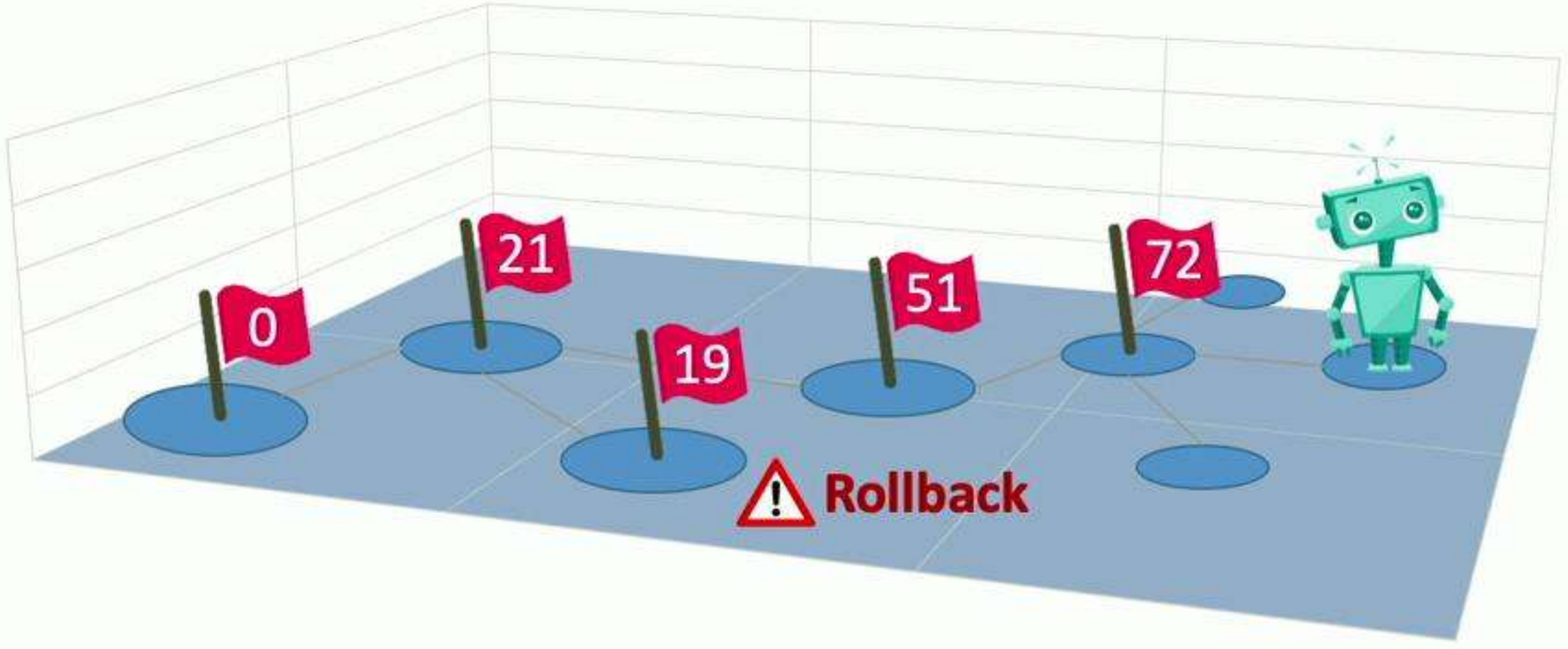


The Regretful Agent – Progress Marker

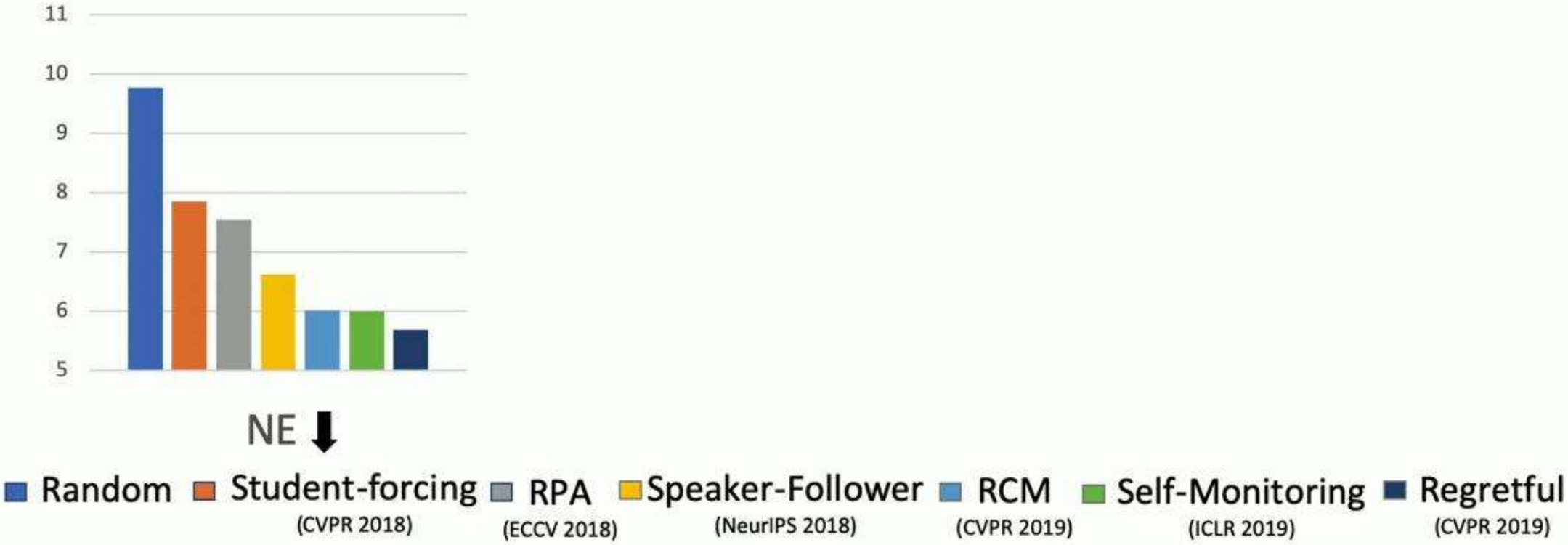
Progress Marker: which way to go?

Local Graph Search:

- know which directions have been visited
- estimate which one is likely to lead to the goal.

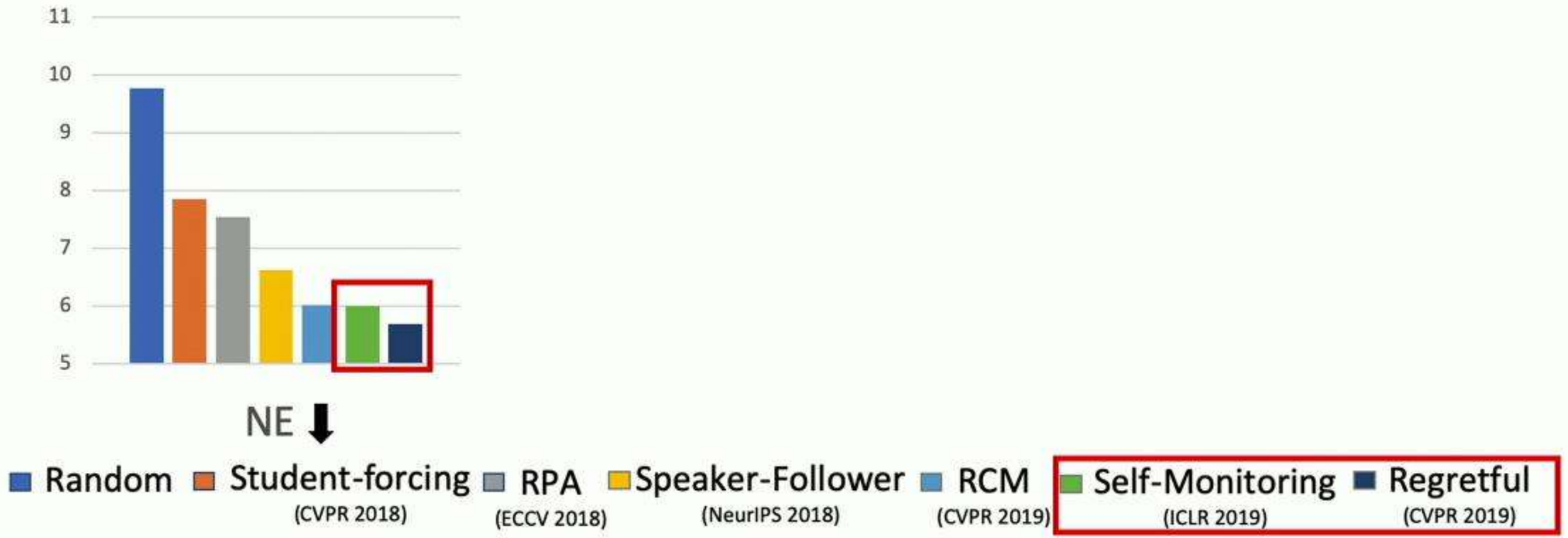


Experiments



Student-forcing (Anderson et al., CVPR 2018), RPA (Wang et al., ECCV 2018), Speaker-Follower (Fried et al., NeurIPS 2018), RCM (Wang et al., CVPR 2019), Self-Monitoring (Ma et al., ICLR 2019), Regretful (Ma et al., CVPR 2019)

Experiments

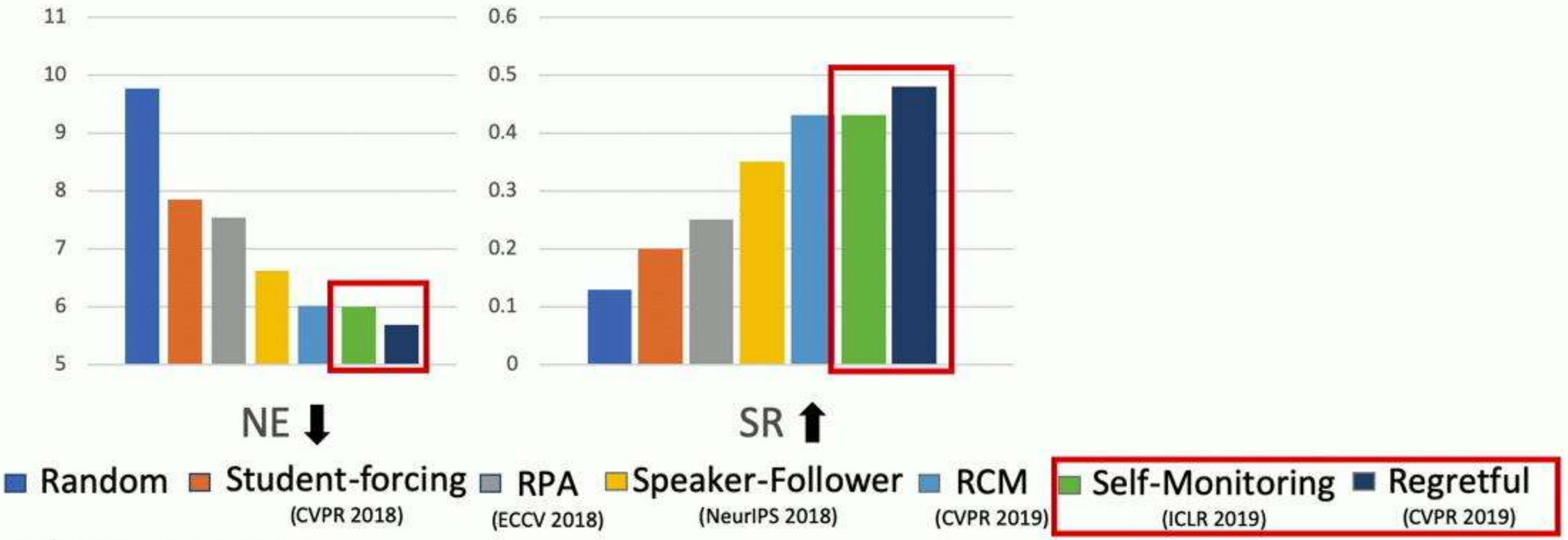


Evaluation metrics:

- Navigation Error (NE): mean of the shortest path distance to the goal location.

Student-forcing (Anderson et al., CVPR 2018), RPA (Wang et al., ECCV 2018), Speaker-Follower (Fried et al., NeurIPS 2018), RCM (Wang et al., CVPR 2019), Self-Monitoring (Ma et al., ICLR 2019), Regretful (Ma et al., CVPR 2019)

Experiments

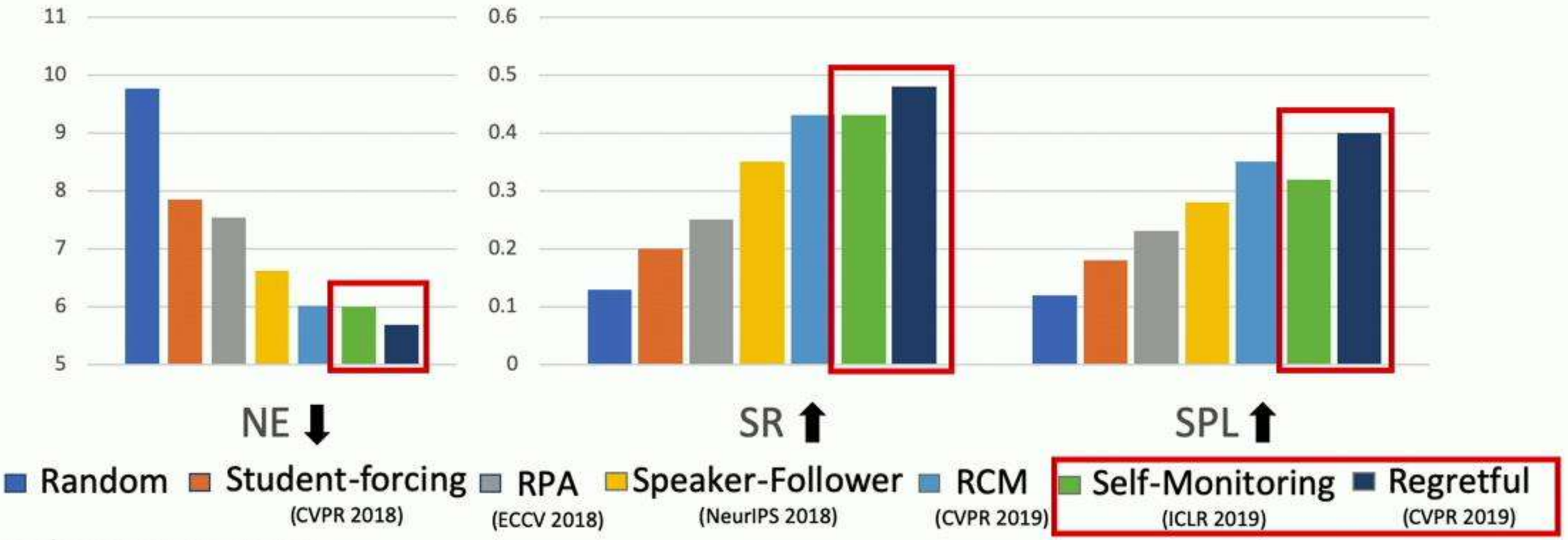


Evaluation metrics:

- Navigation Error (NE): mean of the shortest path distance to the goal location.
- Success Rate (SR): the percentage of final positions less than 3m away from the goal.

Student-forcing (Anderson et al., CVPR 2018), RPA (Wang et al., ECCV 2018), Speaker-Follower (Fried et al., NeurIPS 2018), RCM (Wang et al., CVPR 2019), Self-Monitoring (Ma et al., ICLR 2019), Regretful (Ma et al., CVPR 2019)

Experiments



Evaluation metrics:

- Navigation Error (NE): mean of the shortest path distance to the goal location.
- Success Rate (SR): the percentage of final positions less than 3m away from the goal.
- Success rate weighted by Path Length (SPL): SR weighted with trajectory lengths.

Student-forcing (Anderson et al., CVPR 2018), RPA (Wang et al., ECCV 2018), Speaker-Follower (Fried et al., NeurIPS 2018), RCM (Wang et al., CVPR 2019), Self-Monitoring (Ma et al., ICLR 2019), Regretful (Ma et al., CVPR 2019)

Related Work and Motivations

Grounding issue

Vision-and-Language Navigation

- Anderson et al., CVPR 2018
- Wang et al., ECCV 2018
- Fried et al., NeurIPS 2018

seq-to-seq

soft-attention

RL

⚠ Not properly grounded

Action Recognition

- Simonyan et al., NeurIPS 2014
- Feichtenhofer et al., CVPR 2016
- Sigurdsson et al., CVPR 2016
- Girdhar et al., CVPR 2017
- Carreira et al., CVPR 2017
- Qiu et al., ICCV 2017

LSTM

1D or 3D Conv

CRF

VLAD

optical flow

⚠ No elements to ground

Visual Captioning

- Venugopalan et al., ACL 2014
- Yao et al., ICCV 2015
- Yu et al., CVPR 2016
- Gan et al., CVPR 2017
- Pan et al., CVPR 2017
- Shen et al., CVPR 2017
- Lu et al., CVPR 2018
- Zhou et al., CVPR 2019

seq-to-seq

soft-attention

semantic attribute

RL

⚠ No elements to ground
Grounding rely on supervision

Related Work and Motivations

Grounding issue

Vision-and-Language Navigation

- Anderson et al., CVPR 2018
- Wang et al., ECCV 2018
- Fried et al., NeurIPS 2018

seq-to-seq

soft-attention

RL

⚠ Not properly grounded

Action Recognition

- Simonyan et al., NeurIPS 2014
- Feichtenhofer et al., CVPR 2016
- Sigurdsson et al., CVPR 2016
- Girdhar et al., CVPR 2017
- Carreira et al., CVPR 2017
- Qiu et al., ICCV 2017

LSTM

1D or 3D Conv

CRF

VLAD

optical flow

⚠ No elements to ground

Visual Captioning

- Venugopalan et al., ACL 2014
- Yao et al., ICCV 2015
- Yu et al., CVPR 2016
- Gan et al., CVPR 2017
- Pan et al., CVPR 2017
- Shen et al., CVPR 2017
- Lu et al., CVPR 2018
- Zhou et al., CVPR 2019

seq-to-seq

soft-attention

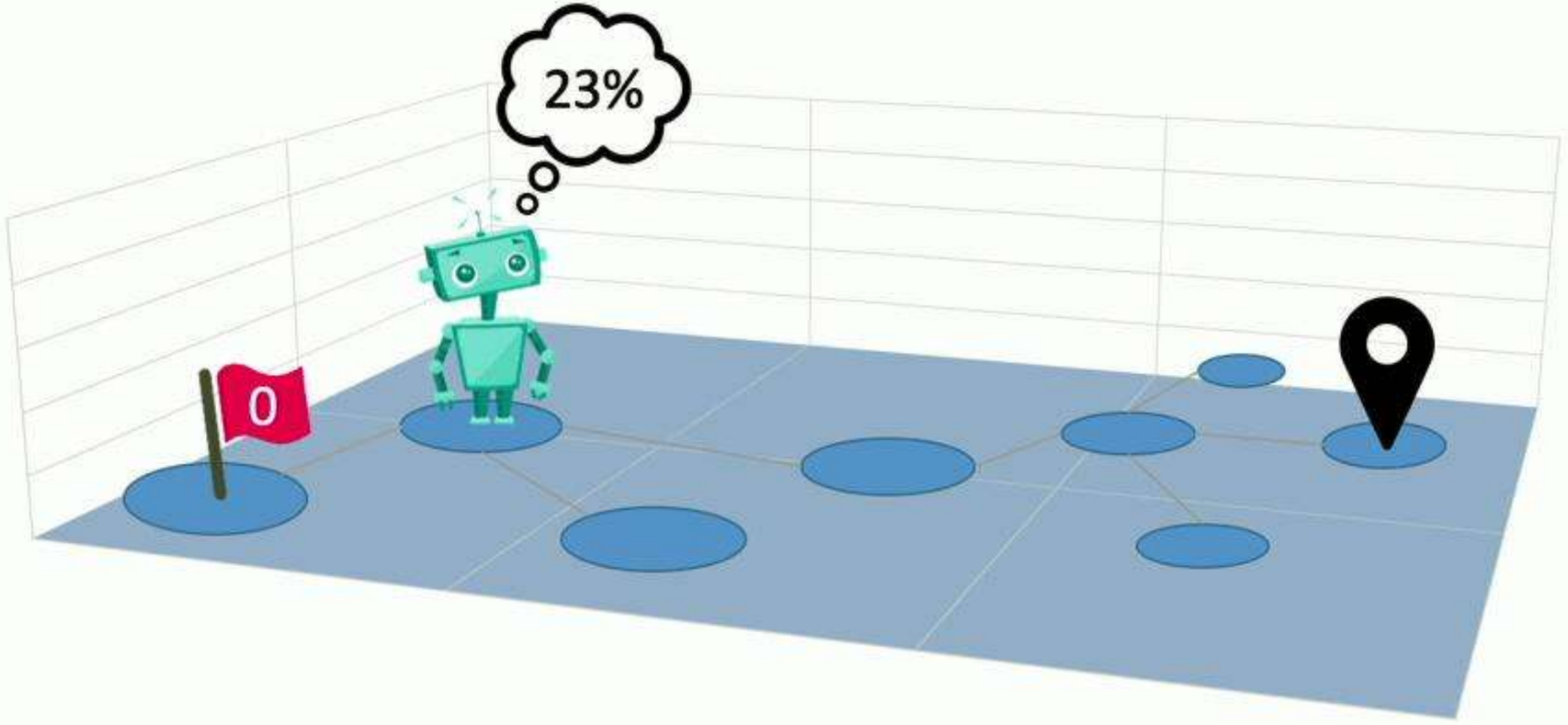
semantic attribute

RL

⚠ No elements to ground
Grounding rely on supervision

The Regretful Agent – Progress Marker

Progress Marker: which way to go?

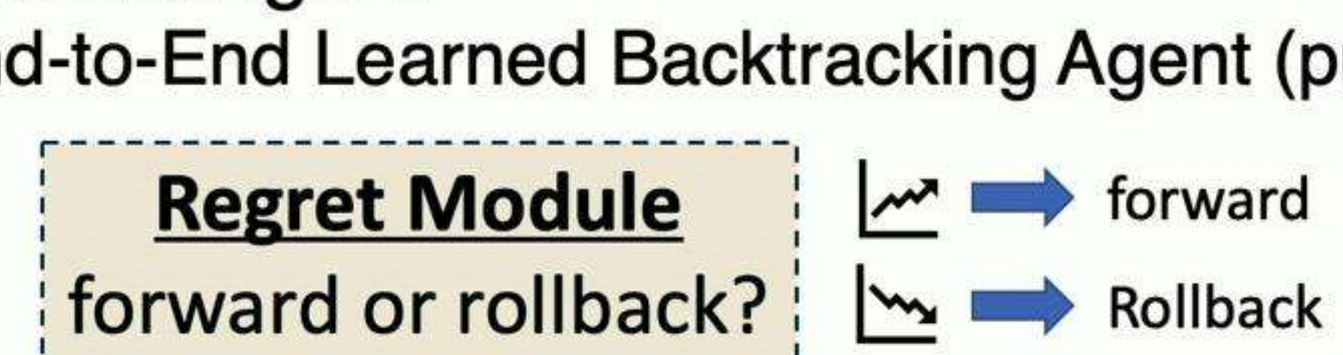


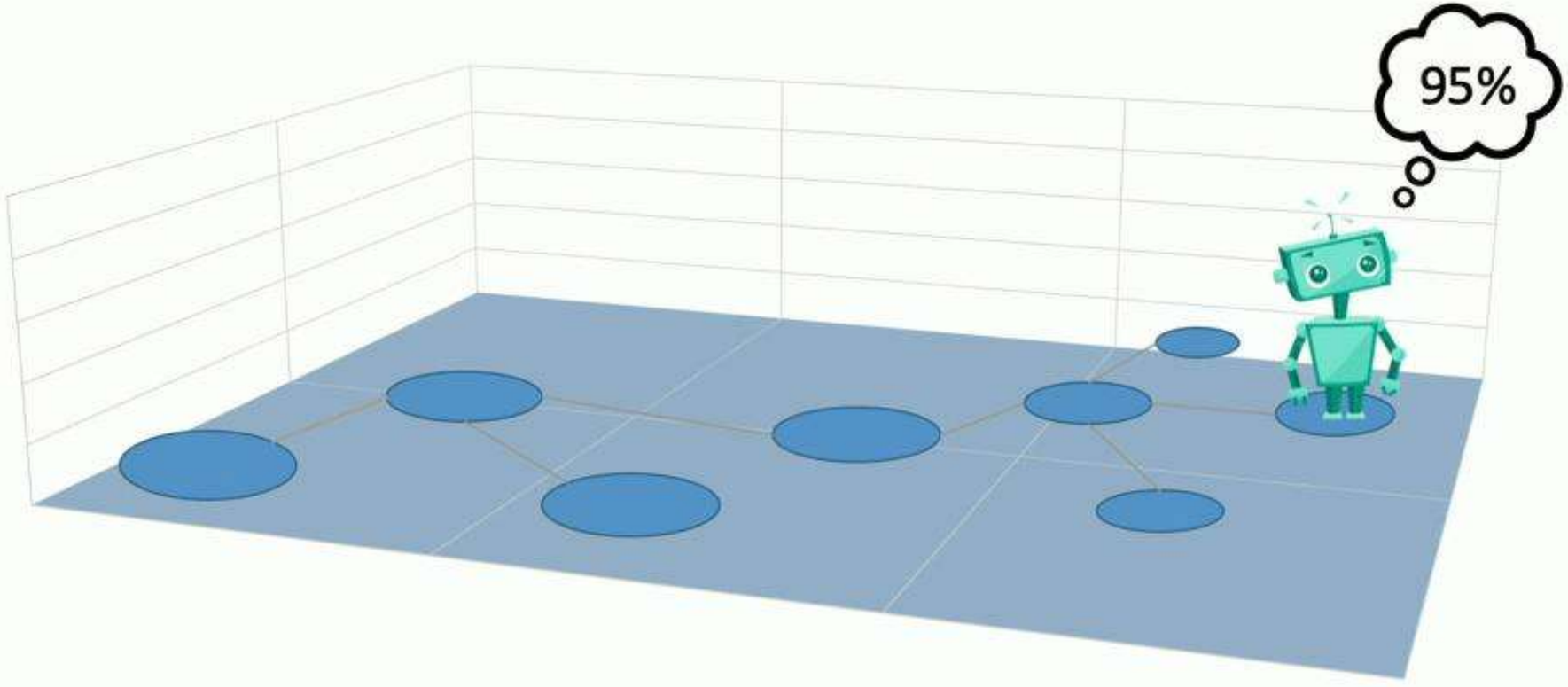
The Regretful Agent

- End-to-End Learned Backtracking Agent (purely greedy)

Regret Module

forward or rollback?





Related Work and Motivations

Grounding issue

Vision-and-Language Navigation

- Anderson et al., CVPR 2018
- Wang et al., ECCV 2018
- Fried et al., NeurIPS 2018

seq-to-seq

soft-attention

RL

⚠ Not properly grounded

Action Recognition

- Simonyan et al., NeurIPS 2014
- Feichtenhofer et al., CVPR 2016
- Sigurdsson et al., CVPR 2016
- Girdhar et al., CVPR 2017
- Carreira et al., CVPR 2017
- Qiu et al., ICCV 2017

LSTM

1D or 3D Conv

CRF

VLAD

optical flow

⚠ No elements to ground

Visual Captioning

- Venugopalan et al., ACL 2014
- Yao et al., ICCV 2015
- Yu et al., CVPR 2016
- Gan et al., CVPR 2017
- Pan et al., CVPR 2017
- Shen et al., CVPR 2017
- Lu et al., CVPR 2018
- Zhou et al., CVPR 2019

seq-to-seq

soft-attention

semantic attribute

RL

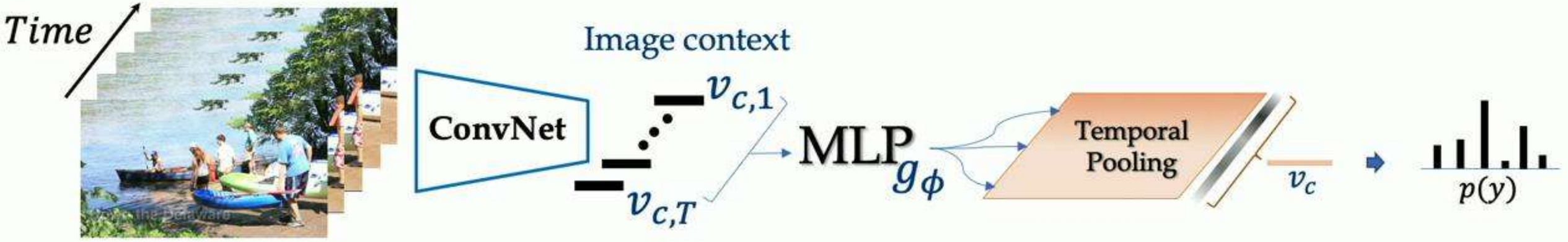
⚠ No elements to ground
Grounding rely on supervision

Outline

- Why do we need grounded spatio-temporal reasoning?
- Related Work & Motivations
- Self-Monitoring and Regretful Navigation Agent
  - Grounding on Vision-and-Language Navigation
- Object Level Fine-Grained Video Understanding
  - Ground human action recognition to object interactions
  - Ground video captioning to object interactions
- Grounded Visual Captioning without human annotations
- Summary

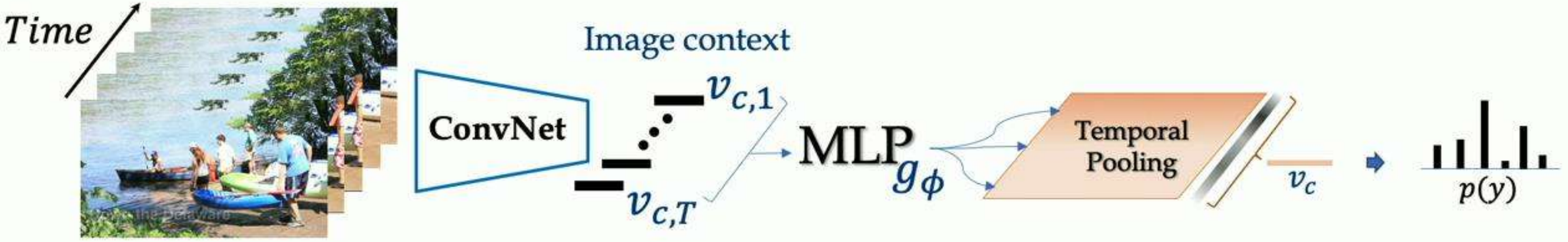
Why Fine-grained Video Understanding?

Existing approaches rely on frame-level representation



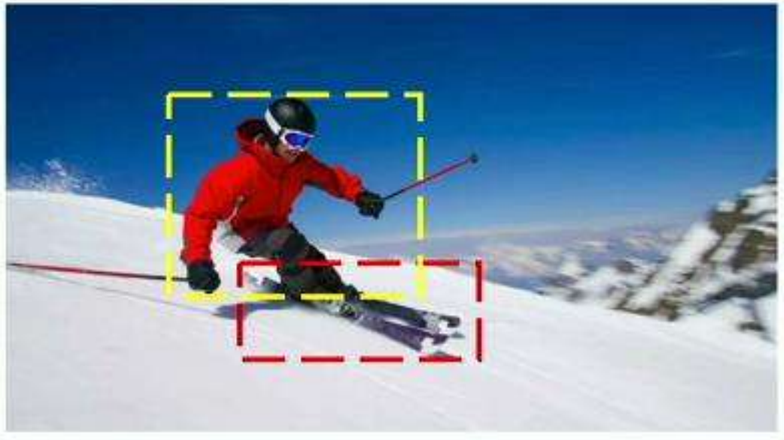
Why Fine-grained Video Understanding?

Existing approaches rely on frame-level representation



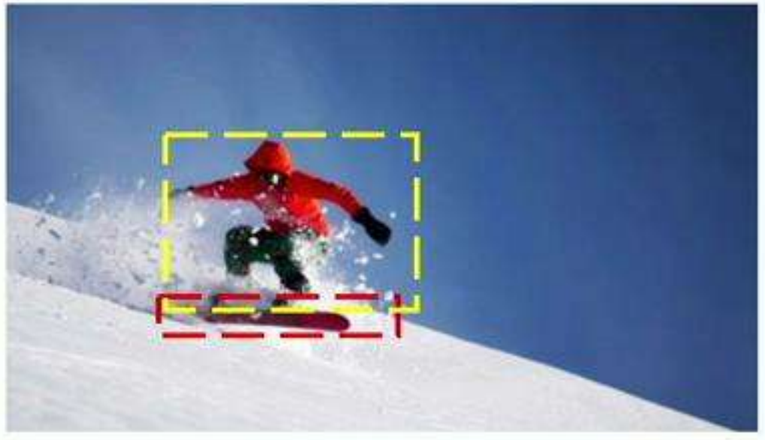
Human action recognition involves complex interactions between objects

Skiing



≠

Snowboarding

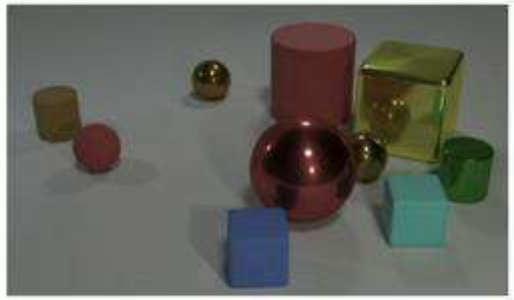


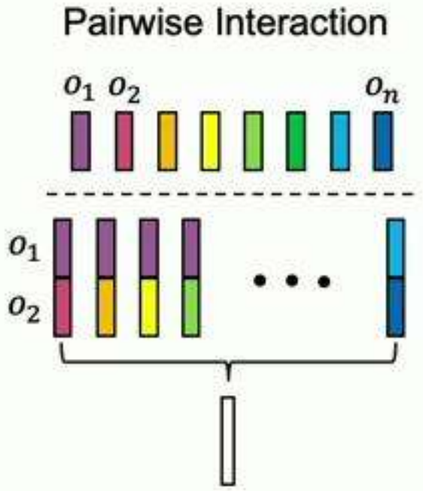
[Person] riding [ski]

[Person] riding [snowboard]

From Dot-Product to Pairwise Relationship

Pairwise relationship [1]:





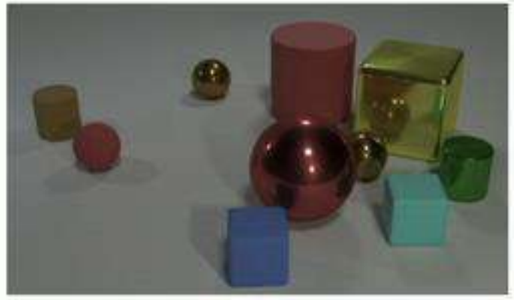
$$RN(O) = f_\phi\left(\sum_{i,j} f_\theta\left(o_i, o_j\right)\right)$$

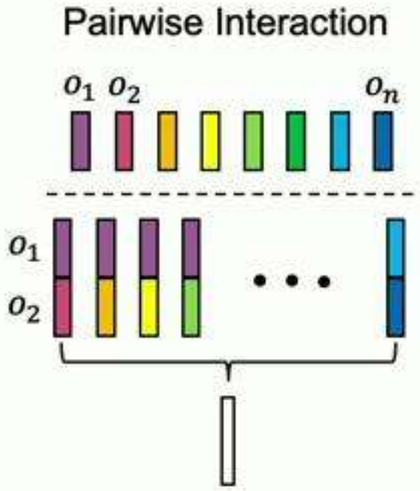
e.g., $W_{f_\theta}^T\left(o_i \,\|\, o_j\right)$

[1] Santoro, Adam, et al. NeurIPS 2017.

From Dot-Product to Pairwise Relationship

Pairwise relationship [1]:





$$RN(O) = f_\phi\left(\sum_{i,j} f_\theta(o_i, o_j)\right)$$

e.g., $W_{f_\theta}^T(o_i \,\|\, o_j)$

Number of objects: 15

$$C_2^{15} = \frac{15\times14}{2} = 105$$

Number of objects: 30

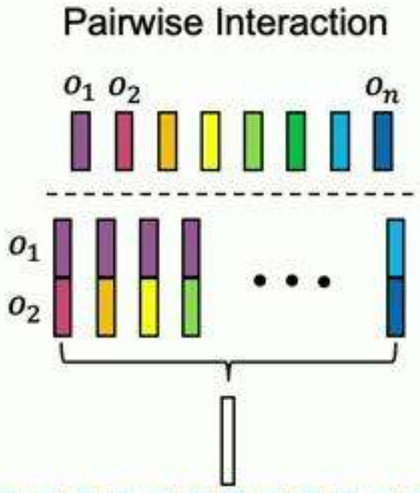
$$C_2^{30} = \frac{30\times29}{2} = 435$$

- Inefficient to detect all relationships across all individual object pairs
- Intractable in video domain

[1] Santoro, Adam, et al. NeurIPS 2017.

From Dot-Product to Pairwise Relationship

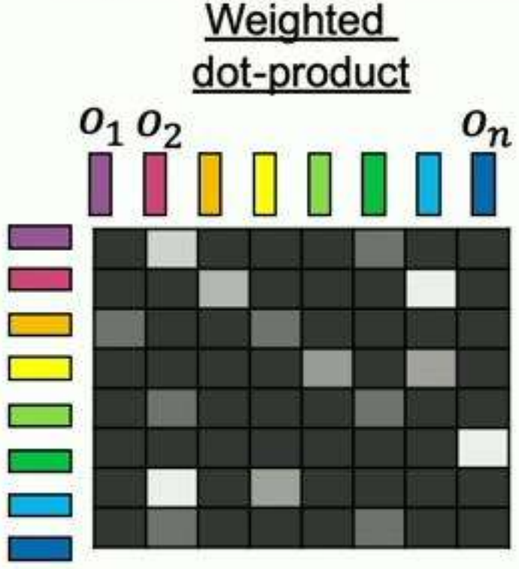
Pairwise relationship [1]:



$$RN(O) = f_\phi\left(\sum_{i,j} f_\theta(o_i, o_j)\right)$$

e.g., $W_{f_\theta}^T(o_i \,\|\, o_j)$

A pairwise function f computes a scalar, representing relationship between i and all j.



$$y_i = \frac{1}{C(o)} \sum_{\forall j} f(o_i, o_j) g(o_j)$$

Normalization

e.g., $W_g o_j$

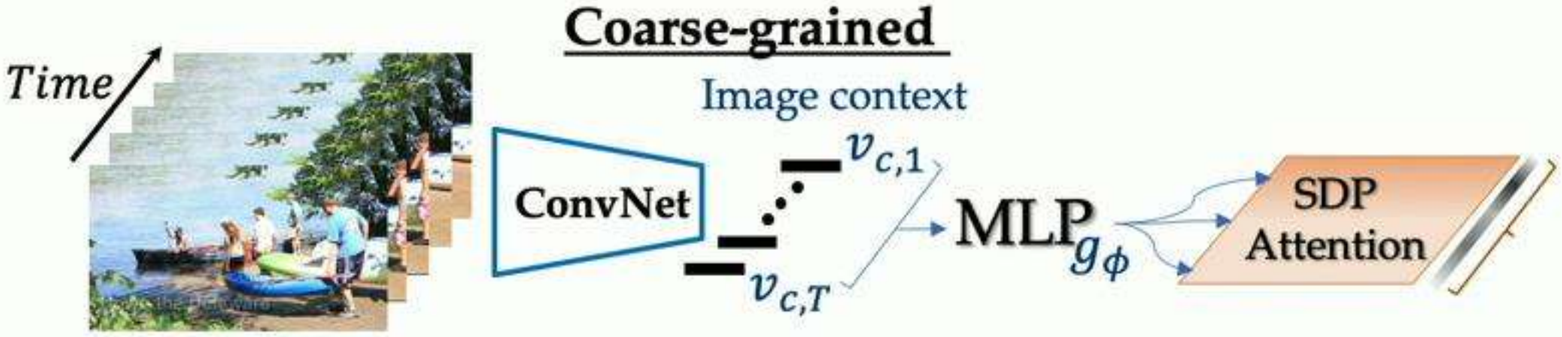
Weighted dot-product:

$$f(o_i, o_j) = \theta(o_i)^T \phi(o_j)$$

$$= O^T W_\theta^T W_\phi O$$

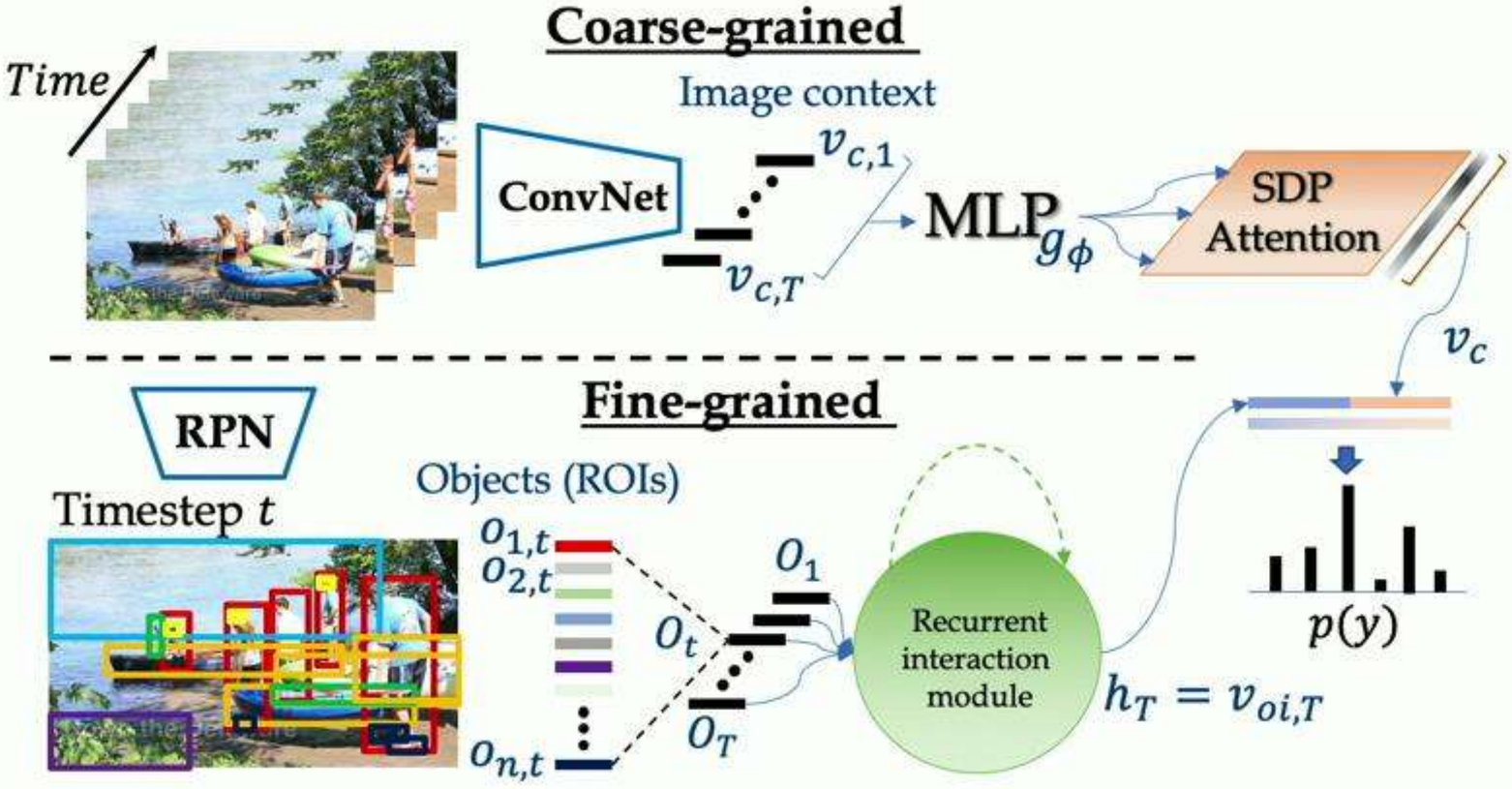
From Object Interaction to Human Action Recognition

- **Coarse-grained**: each video frame is encoded into a feature vector. The sequence of vectors are then pooled via temporal SDP-Attention into single vector representation.



From Object Interaction to Human Action Recognition

- **Coarse-grained**: each video frame is encoded into a feature vector. The sequence of vectors are then pooled via temporal SDP-Attention into single vector representation.
- **Fine-grained**: each object (ROI) obtained from RPN is encoded in a feature vector. We detect the higher-order object interaction using the proposed generic recurrent interaction module.

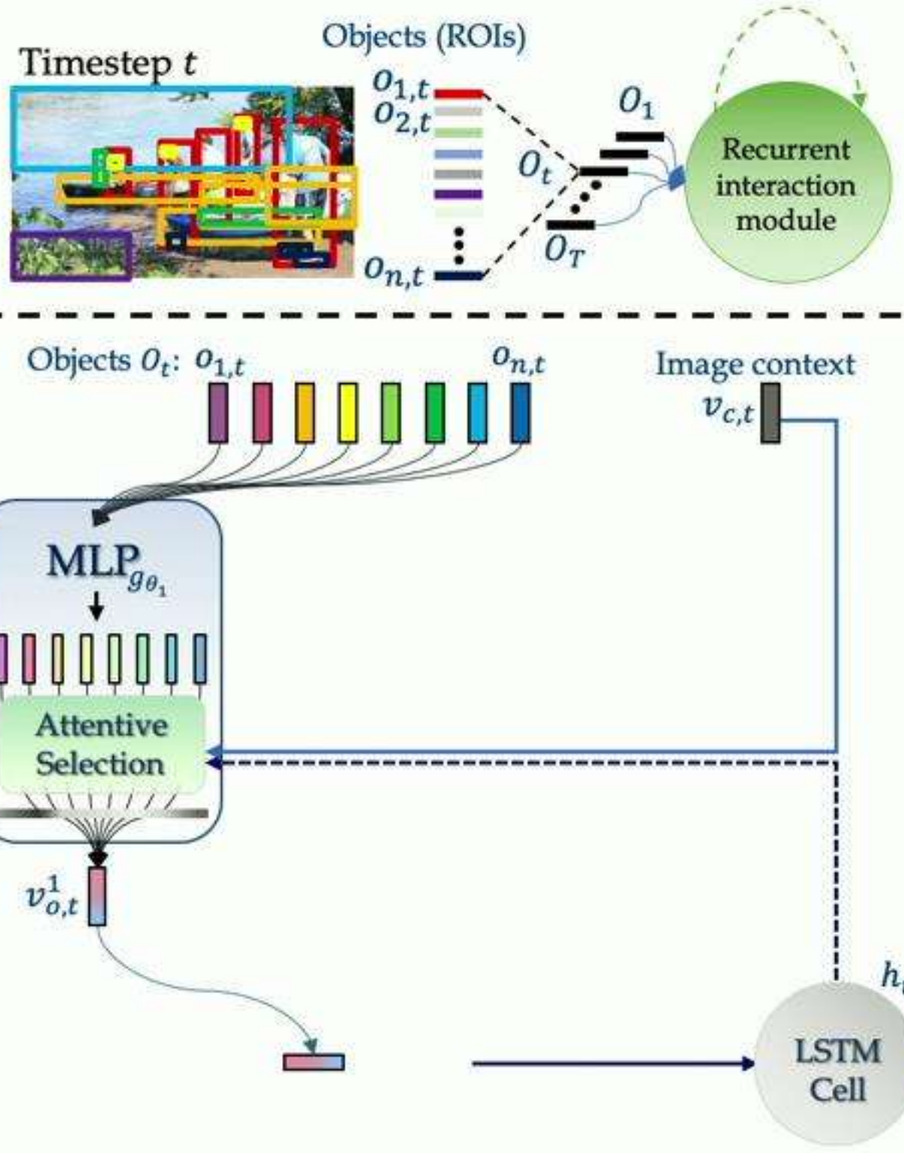


Recurrent Interaction Module

- Dynamically select objects which are important to discriminate human actions via Dot-Product Attention.

$$\alpha_k = Attention\left(g_{\theta_k}(O_t), v_{c,t}, h_{t-1}\right)$$

Object features | Image feature | Previous interaction

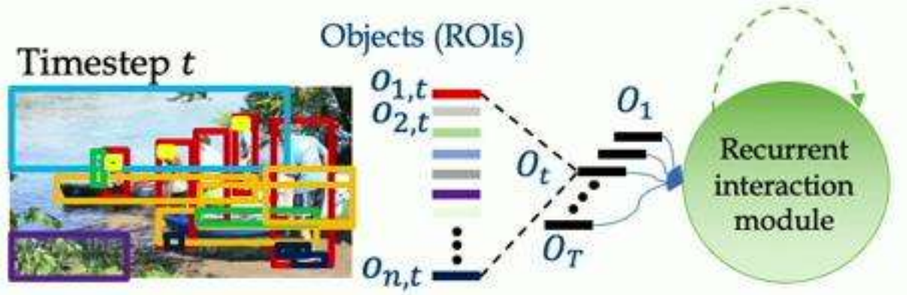


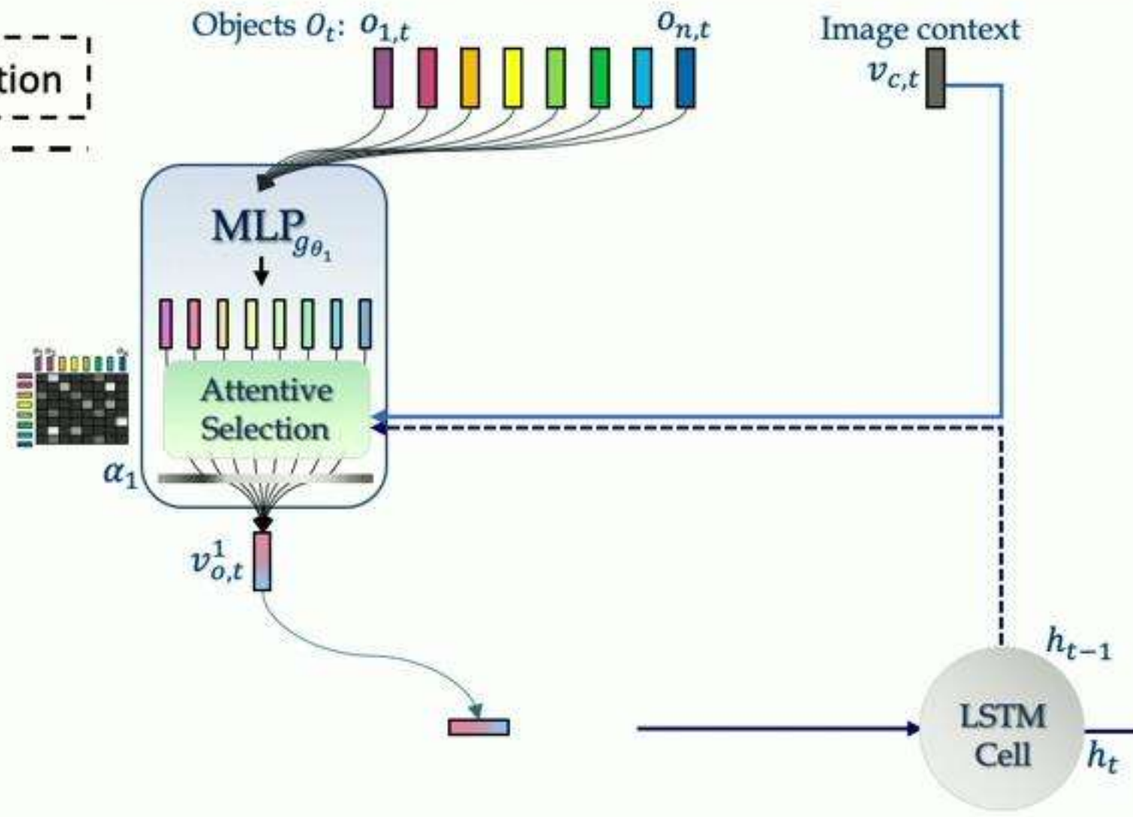
Recurrent Interaction Module

- Dynamically select objects which are important to discriminate human actions via Dot-Product Attention.

$$\alpha_k = Attention\left(g_{\theta_k}(O_t), v_{c,t}, h_{t-1}\right)$$

Object features | Image feature | Previous interaction





Group to group

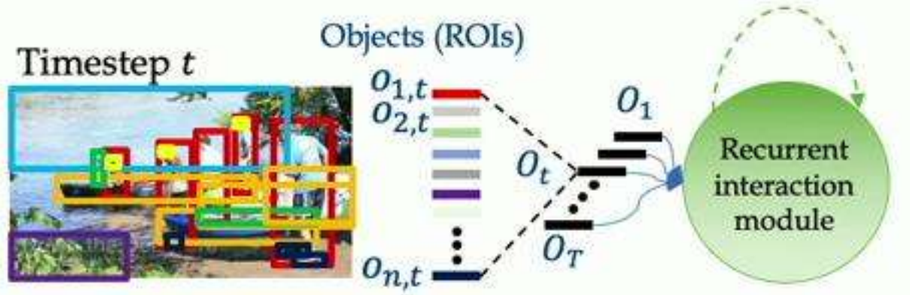
[Man with glasses] talking [Man on a chair]

Recurrent Interaction Module

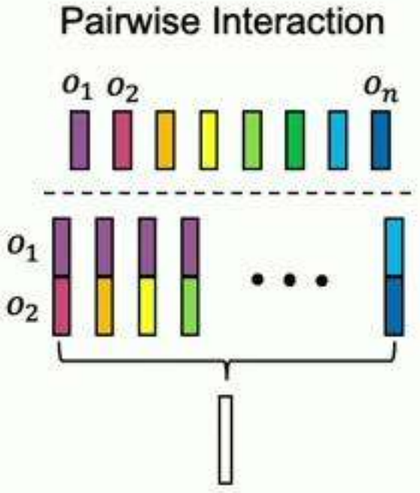
- Dynamically select objects which are important to discriminate human actions via Dot-Product Attention.

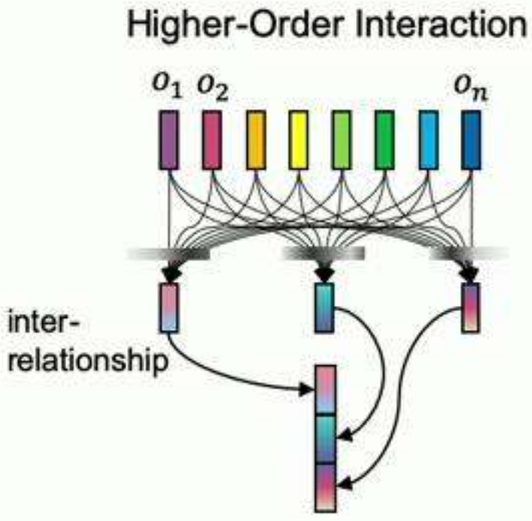
$$\alpha_k = Attention\left(g_{\theta_k}(O_t), v_{c,t}, h_{t-1}\right)$$

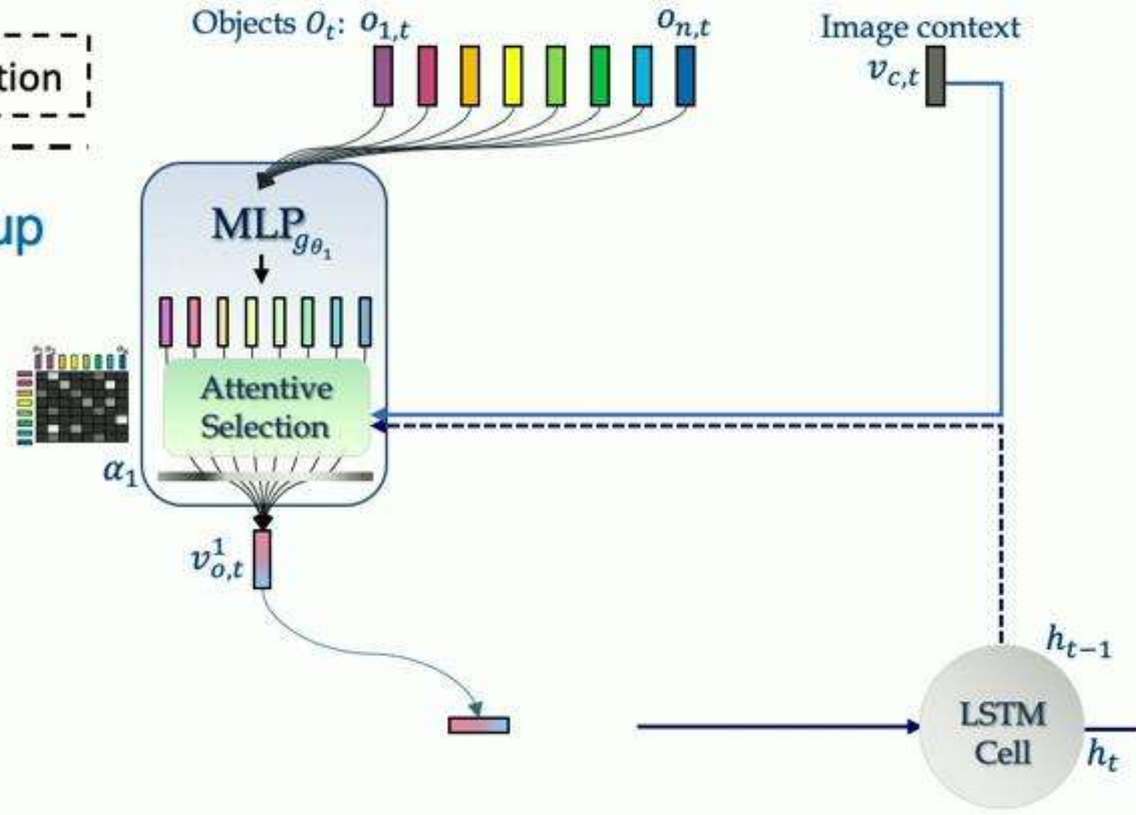
Object features | Image feature | Previous interaction



- Model higher-order interaction using group to group or triplet groups of objects via concatenation.





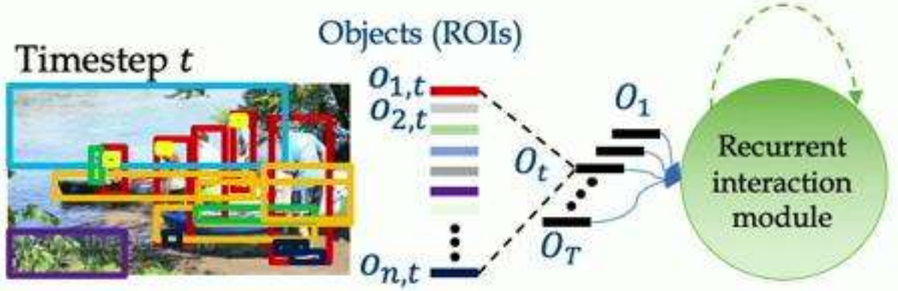


Recurrent Interaction Module

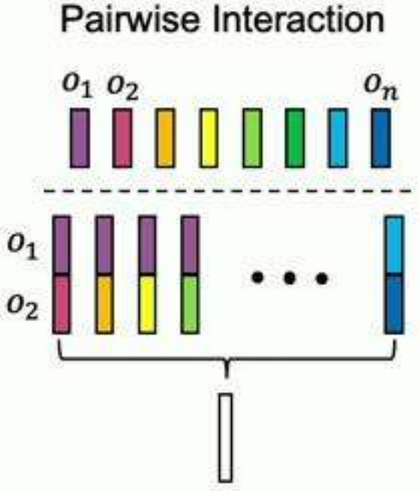
- Dynamically select objects which are important to discriminate human actions via Dot-Product Attention.

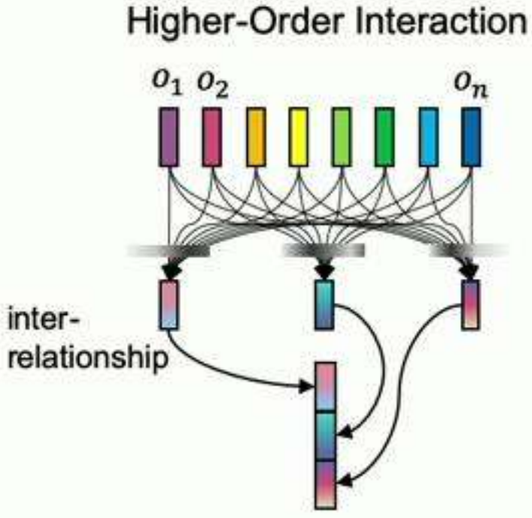
$$\alpha_k = Attention\left(g_{\theta_k}(O_t), v_{c,t}, h_{t-1}\right)$$

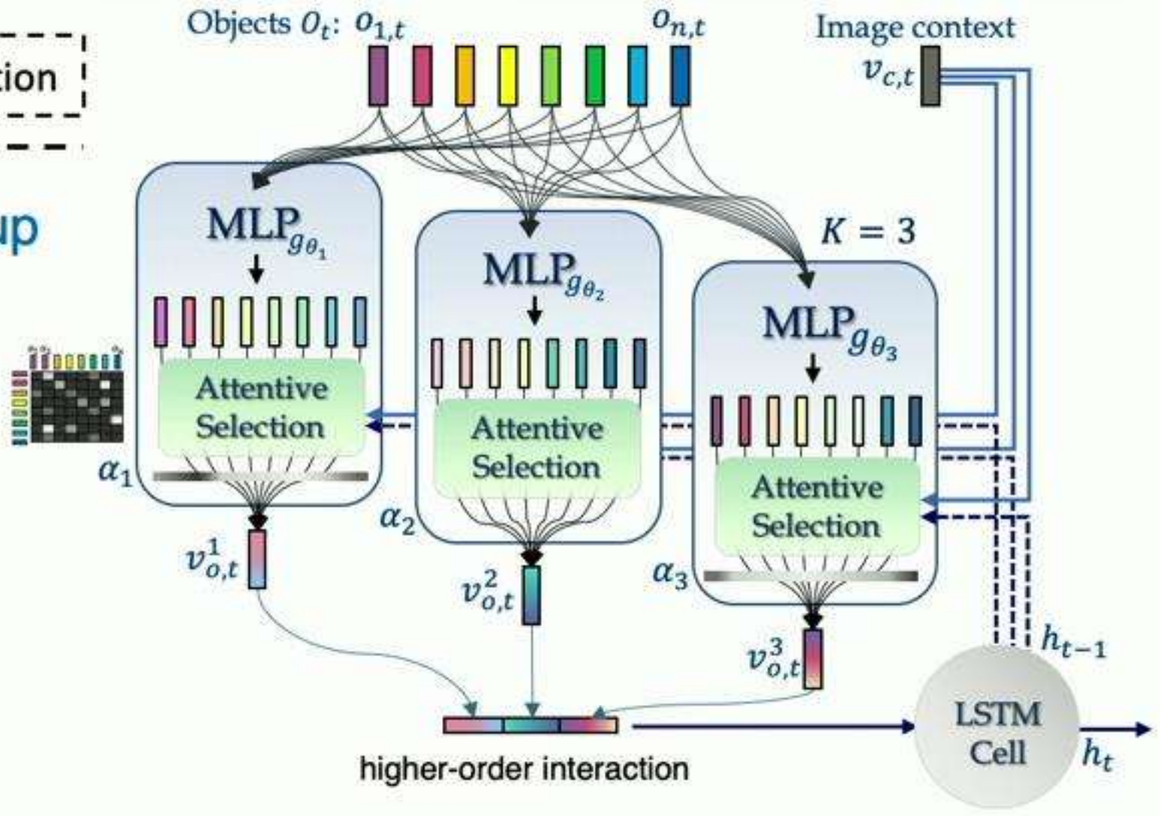
Object features | Image feature | Previous interaction



- Model higher-order interaction using group to group or triplet groups of objects via concatenation.







Experiments on Kinetics

- Kinetics dataset:
  - 400 human action classes
  - 300,000 videos (833 video hours)
  - Each video is 10 seconds long

Method	Top-1	Top-5
I3D (25 FPS) (test)	71.1	89.3
TSN (Inception-ResNet-v2) (2.5 FPS)	73.0	90.9
Ours (1 FPS)		
Img feature + LSTM (baseline)	70.6	89.1
Img feature + temporal SDP-Attention	71.1	89.6
SINet (dot-product attention)	**74.2**	**91.7**

I3D (Carreira, et al., CVPR 2017), TSN (Wang et al., ECCV 2016), SINet (Ma et al., CVPR 2018)

Experiments on Kinetics

- Kinetics dataset:
  - 400 human action classes
  - 300,000 videos (833 video hours)
  - Each video is 10 seconds long

- Outperformed existing state of the arts with lower video sampling rate.

Method	Top-1	Top-5
I3D (25 FPS) (test)	71.1	89.3
TSN (Inception-ResNet-v2) (2.5 FPS)	73.0	90.9
Ours (1 FPS)		
Img feature + LSTM (baseline)	70.6	89.1
Img feature + temporal SDP-Attention	71.1	89.6
SINet (dot-product attention)	**74.2**	**91.7**

I3D (Carreira, et al., CVPR 2017), TSN (Wang et al., ECCV 2016), SINet (Ma et al., CVPR 2018)

Experiments on Kinetics

- Kinetics dataset:
  - 400 human action classes
  - 300,000 videos (833 video hours)
  - Each video is 10 seconds long

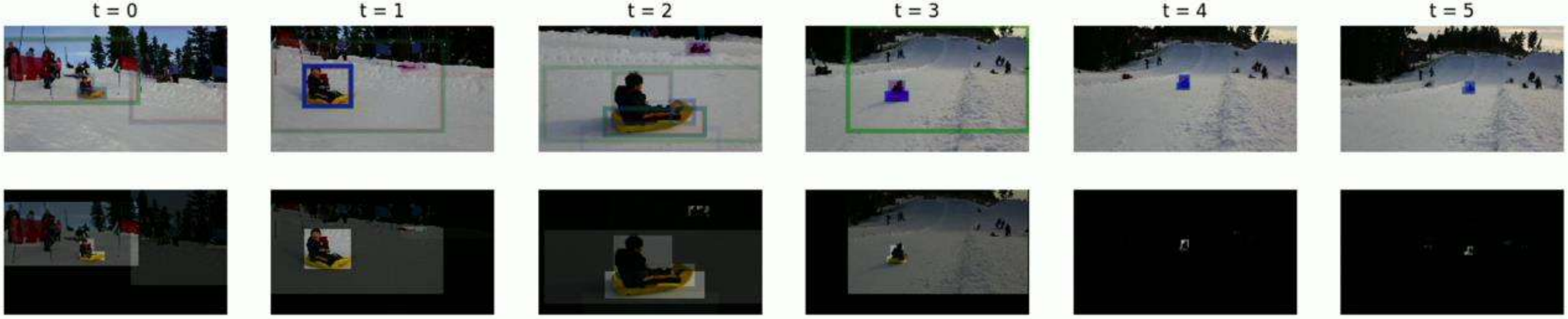
Method	Top-1	Top-5
I3D (25 FPS) (test)	71.1	89.3
TSN (Inception-ResNet-v2) (2.5 FPS)	73.0	90.9
Ours (1 FPS)		
Img feature + LSTM (baseline)	70.6	89.1
Img feature + temporal SDP-Attention	71.1	89.6
SINet (dot-product attention)	**74.2**	**91.7**

Method	Top-1	Top-5	FLOP (e^9)
Obj (mean-pooling)	73.1	90.8	1.9
Obj pairs (mean-pooling)	73.4	90.8	18.3
Obj triplet (mean-pooling)	72.9	90.7	77.0
SINet ($K = 1$)	73.9	91.3	2.7
SINet ($K = 2$)	74.2	91.5	5.3
SINet ($K = 3$)	**74.2**	**91.7**	8.0

- Outperformed existing state of the arts with lower video sampling rate.
- Modeling higher-order interactions efficiently.

I3D (Carreira, et al., CVPR 2017), TSN (Wang et al., ECCV 2016), SINet (Ma et al., CVPR 2018)

Qualitative Analysis on Kinetics



Tobogganing

1. *Toboggan*
2. *snow scene*
3. *person*

- Successfully detect three key objects and their relationships
- Track the person and toboggan
- Completely ignore the irrelevant background afterwards

Experiments on Kinetics

- Kinetics dataset:
  - 400 human action classes
  - 300,000 videos (833 video hours)
  - Each video is 10 seconds long

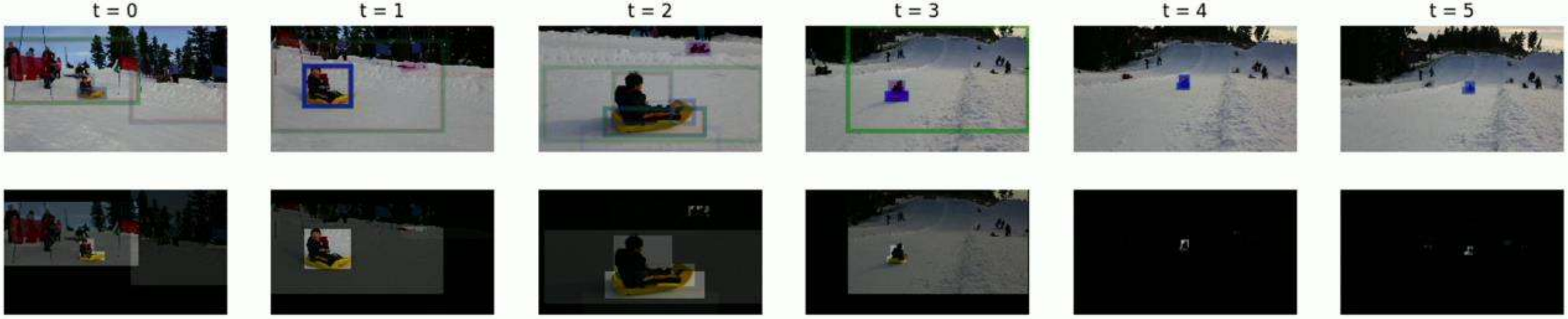
- Outperformed existing state of the arts with lower video sampling rate.
- Modeling higher-order interactions efficiently.

Method	Top-1	Top-5
I3D (25 FPS) (test)	71.1	89.3
TSN (Inception-ResNet-v2) (2.5 FPS)	73.0	90.9
Ours (1 FPS)		
Img feature + LSTM (baseline)	70.6	89.1
Img feature + temporal SDP-Attention	71.1	89.6
SINet (dot-product attention)	**74.2**	**91.7**

Method	Top-1	Top-5	FLOP (e^9)
Obj (mean-pooling)	73.1	90.8	1.9
Obj pairs (mean-pooling)	73.4	90.8	18.3
Obj triplet (mean-pooling)	72.9	90.7	77.0
SINet ($K = 1$)	73.9	91.3	2.7
SINet ($K = 2$)	74.2	91.5	5.3
SINet ($K = 3$)	**74.2**	**91.7**	8.0

I3D (Carreira, et al., CVPR 2017), TSN (Wang et al., ECCV 2016), SINet (Ma et al., CVPR 2018)

Qualitative Analysis on Kinetics



Tobogganing

1. *Toboggan*
2. *snow scene*
3. *person*

- Successfully detect three key objects and their relationships
- Track the person and toboggan
- Completely ignore the irrelevant background afterwards

Experiments on Kinetics

- Kinetics dataset:
  - 400 human action classes
  - 300,000 videos (833 video hours)
  - Each video is 10 seconds long

- Outperformed existing state of the arts with lower video sampling rate.
- Modeling higher-order interactions efficiently.

Method	Top-1	Top-5
I3D (25 FPS) (test)	71.1	89.3
TSN (Inception-ResNet-v2) (2.5 FPS)	73.0	90.9
Ours (1 FPS)		
Img feature + LSTM (baseline)	70.6	89.1
Img feature + temporal SDP-Attention	71.1	89.6
SINet (dot-product attention)	**74.2**	**91.7**

Method	Top-1	Top-5	FLOP (e^9)
Obj (mean-pooling)	73.1	90.8	1.9
Obj pairs (mean-pooling)	73.4	90.8	18.3
Obj triplet (mean-pooling)	72.9	90.7	77.0
SINet ($K = 1$)	73.9	91.3	2.7
SINet ($K = 2$)	74.2	91.5	5.3
SINet ($K = 3$)	**74.2**	**91.7**	8.0

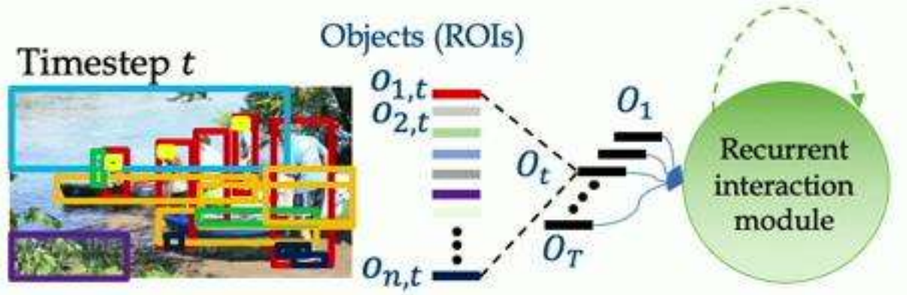
I3D (Carreira, et al., CVPR 2017), TSN (Wang et al., ECCV 2016), SINet (Ma et al., CVPR 2018)

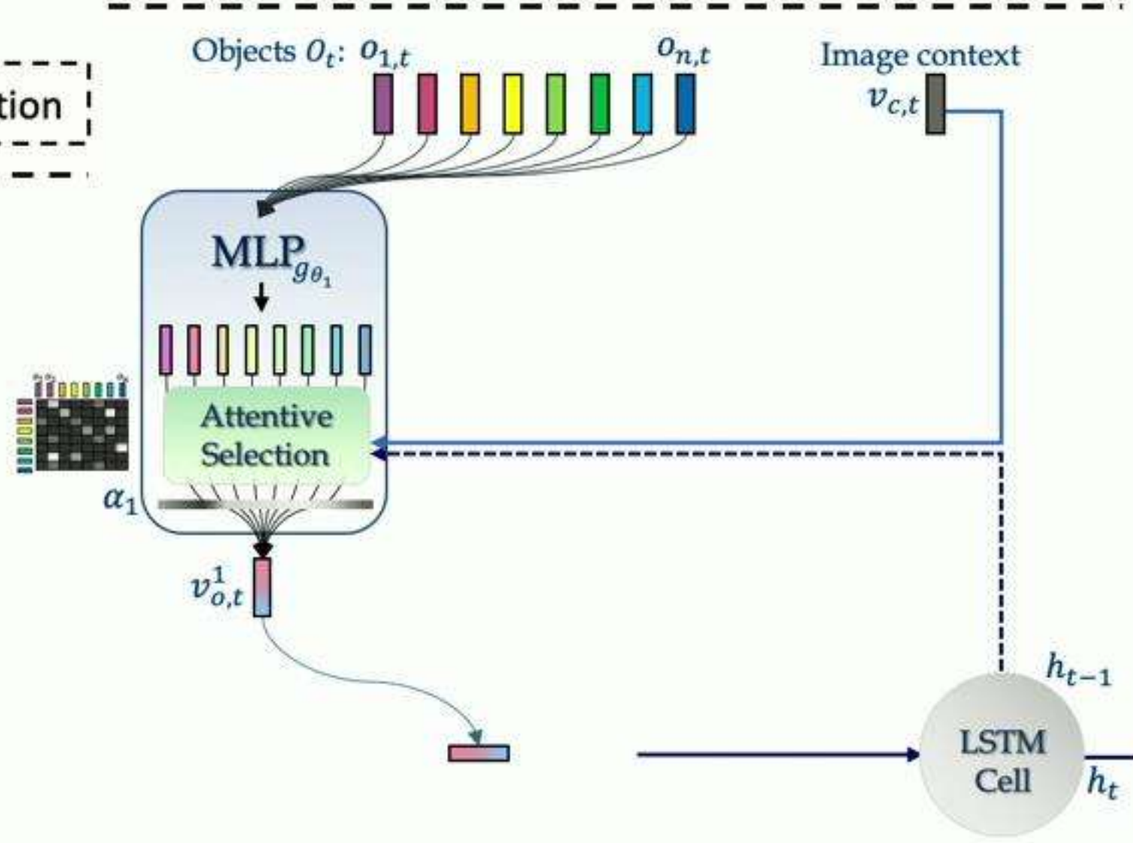
Recurrent Interaction Module

- Dynamically select objects which are important to discriminate human actions via Dot-Product Attention.

$$\alpha_k = Attention\left(g_{\theta_k}(O_t), v_{c,t}, h_{t-1}\right)$$

Object features | Image feature | Previous interaction





Group to group

[Man with glasses]

talking

[Man on a chair]

Experiments on Kinetics

- Kinetics dataset:
  - 400 human action classes
  - 300,000 videos (833 video hours)
  - Each video is 10 seconds long

- Outperformed existing state of the arts with lower video sampling rate.

Method	Top-1	Top-5
I3D (25 FPS) (test)	71.1	89.3
TSN (Inception-ResNet-v2) (2.5 FPS)	73.0	90.9
Ours (1 FPS)		
Img feature + LSTM (baseline)	70.6	89.1
Img feature + temporal SDP-Attention	71.1	89.6
SINet (dot-product attention)	**74.2**	**91.7**

I3D (Carreira, et al., CVPR 2017), TSN (Wang et al., ECCV 2016), SINet (Ma et al., CVPR 2018)

From Object Interaction to Video Captioning

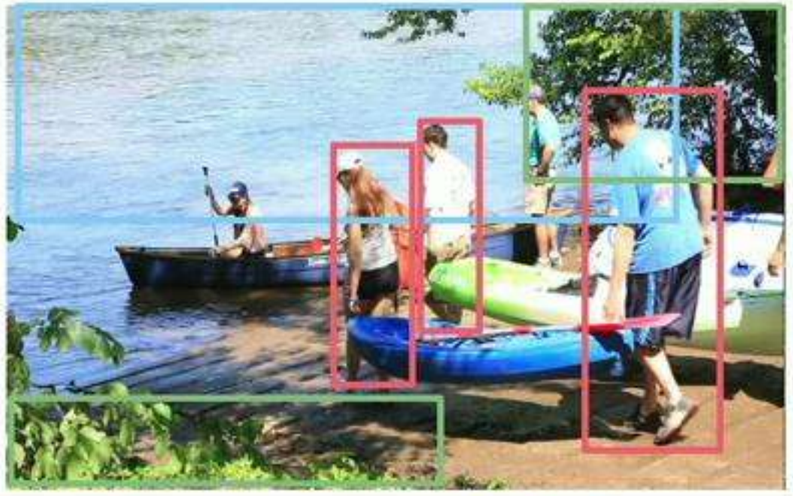
A group of people get off of a yellow school bus with life vest around their neck and enter a foliage filled area leading to a body of water.

Relationships: [group of people, get off, yellow school bus], [group of people, with, life vest], [group of people, enter, foliage filled area], [foliage filled area, leading to, body of water]

Insight:

A caption can almost always be decomposed into a set of object relationship/interactions.

From Object Interaction to Video Captioning



A group of people get off of a yellow school bus with life vest around their neck and enter a foliage filled area leading to a body of water.

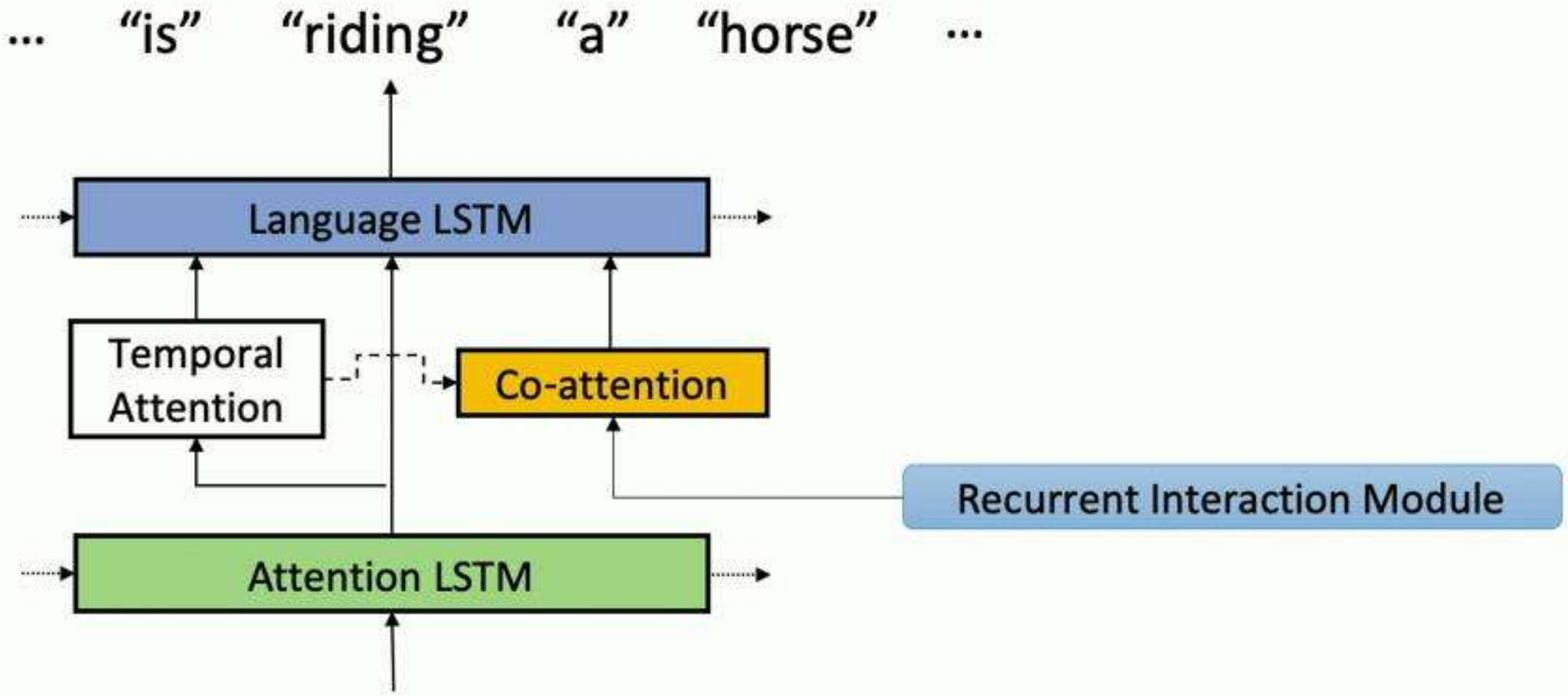
Relationships: [group of people, get off, yellow school bus], [group of people, with, life vest], [group of people, enter, foliage filled area], [foliage filled area, leading to, body of water]

Insight:

A caption can almost always be decomposed into a set of object relationship/interactions.

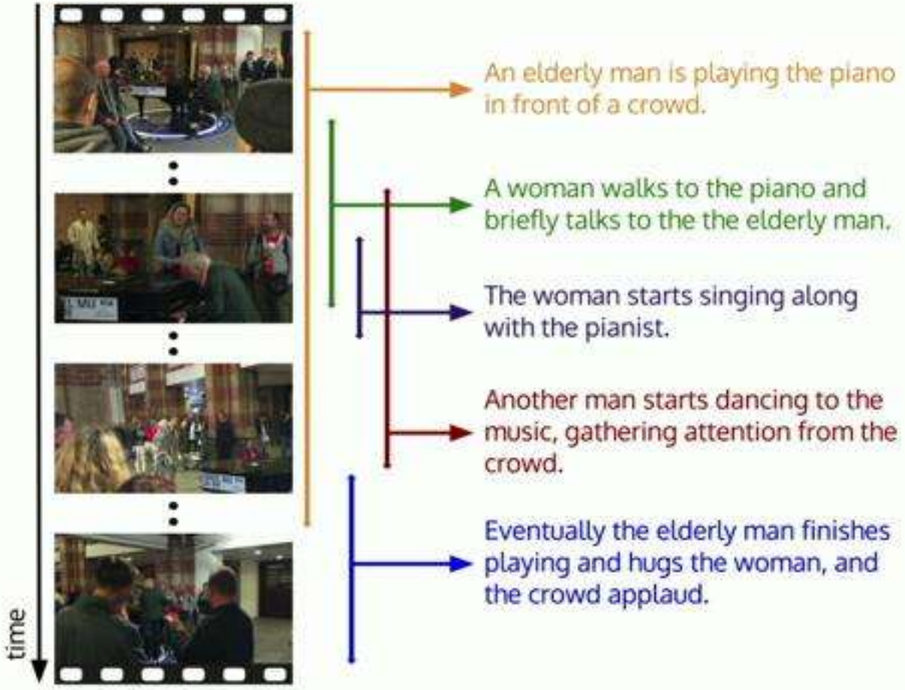
How do we compose a description using a set of detected object relationship/interactions?

Video Captioning – SINet-Caption



- The Attention LSTM identifies which part of the video in spatiotemporal feature space is needed for Language LSTM to generate the next word.

Experiments on ActivityNet Caption



Captioning Evaluation Metrics

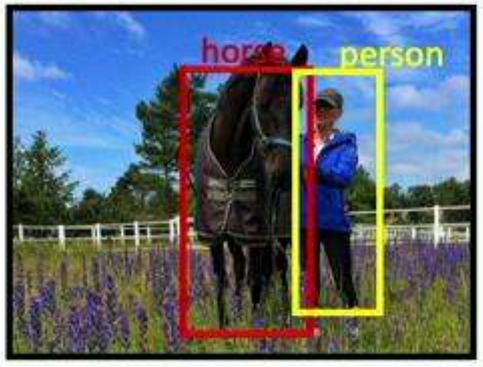
Method	B@1	B@2	B@3	B@4	ROUGE-L	METEOR	CIDEr-D
Test set							
LSTM-YT (C3D)	18.22	7.43	3.24	1.24	-	6.56	14.86
S2VT (C3D)	20.35	8.99	4.60	2.62	-	7.85	20.97
H-RNN (C3D)	19.46	8.78	4.34	2.53	-	8.02	20.18
S2VT + full context (C3D)	26.45	13.48	7.21	3.98	-	9.46	24.56
LSTM-A_3 + policy gradient + retrieval (ResNet + P3D ResNet)	-	-	-	-	-	12.84	-
Validation set (Avg. 1st and 2nd)							
LSTM-A_3 (ResNet + P3D ResNet)	17.5	9.62	5.54	**3.38**	13.27	7.71	16.08
LSTM-A_3 + policy gradient + retrieval (ResNet + P3D ResNet)	17.27	9.70	5.39	3.13	14.29	8.73	14.75
SINet-Caption (ResNeXt)	**19.78**	**9.89**	**4.52**	1.98	**21.25**	**9.84**	**44.84**

- ActivityNet Captions
  - 20,000 videos (849 video hours)
  - 3.65 temporal segments
  - 100,000 captions
  - Average length is 180 seconds

- We compare with state of the arts on validation sets and indirectly on the test set.
- We significantly outperformed existing approaches with large margin.

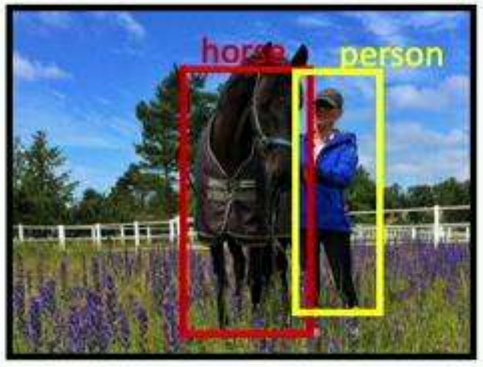
LSTM-YT (Venugopalan et al., ACL 2015), S2VT (Venugopalan et al., ICCV 2015), H-RNN (Yu et al., CVPR 2016), LSTM-A3 (Yao et al., arXiv 2018), SINet-Caption (Ma et al., CVPR 2018)

Recognize Actions When Common Objects Presented



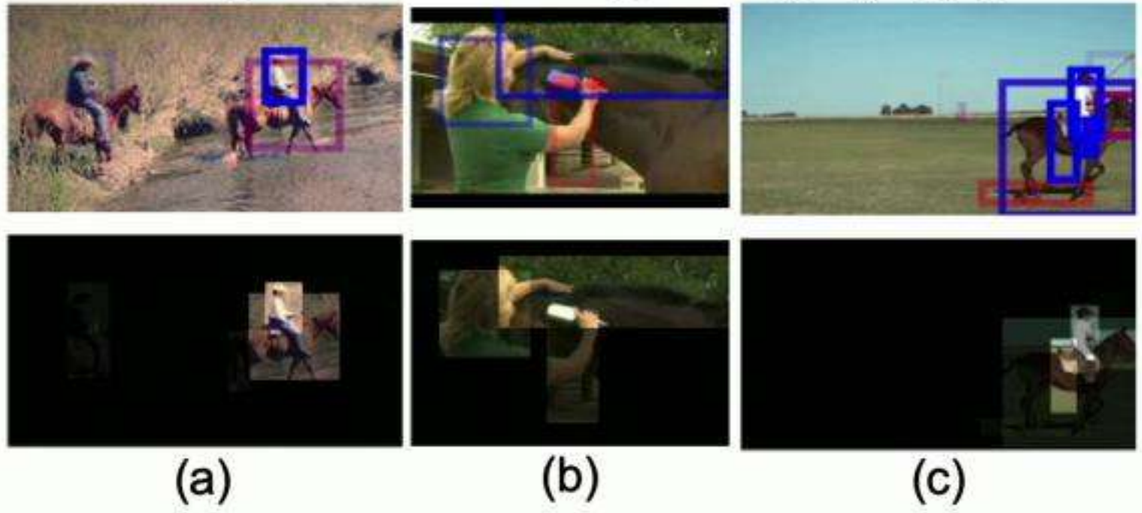
A man is riding on a horse!

Recognize Actions When Common Objects Presented



A man is riding on a horse!

- Distinguishes the interactions of the human and the horse.
- Each video shares a common object in the scene – **horse**
  a) People are *riding* horses.
  b) A woman is *brushing* a horse.
  c) People are *playing polo* on a field.



Related Work and Motivations

Grounding issue

Vision-and-Language Navigation

- Anderson et al., CVPR 2018
- Wang et al., ECCV 2018
- Fried et al., NeurIPS 2018

seq-to-seq

soft-attention

RL

⚠ Not properly grounded

✓

Action Recognition

- Simonyan et al., NeurIPS 2014
- Feichtenhofer et al., CVPR 2016
- Sigurdsson et al., CVPR 2016
- Girdhar et al., CVPR 2017
- Carreira et al., CVPR 2017
- Qiu et al., ICCV 2017

LSTM

1D or 3D Conv

CRF

VLAD

optical flow

⚠ No elements to ground

✓

Visual Captioning

- Venugopalan et al., ACL 2014
- Yao et al., ICCV 2015
- Yu et al., CVPR 2016
- Gan et al., CVPR 2017
- Pan et al., CVPR 2017
- Shen et al., CVPR 2017
- Lu et al., CVPR 2018
- Zhou et al., CVPR 2019

seq-to-seq

soft-attention

semantic attribute

RL

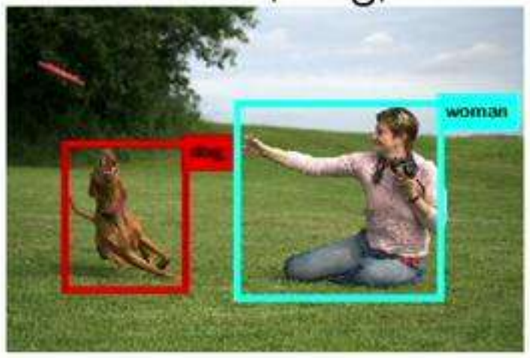
⚠ No elements to ground ✓

Grounding rely on supervision

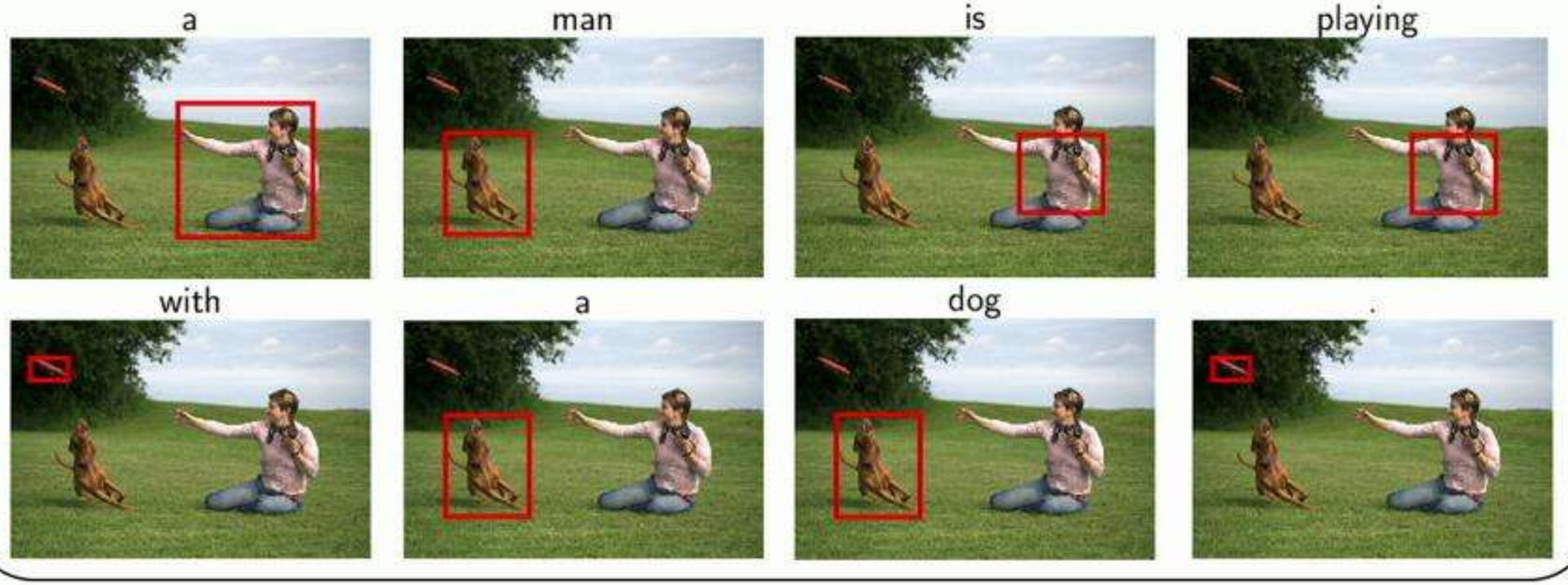
Outline

- Why do we need grounded spatio-temporal reasoning?
- Related Work & Motivations
- Self-Monitoring and Regretful Navigation Agent
  - Grounding on Vision-and-Language Navigation
- Object Level Fine-Grained Video Understanding
  - Ground human action recognition to object interactions
  - Ground video captioning to object interactions
- **Grounded Visual Captioning without human annotations**
- Summary

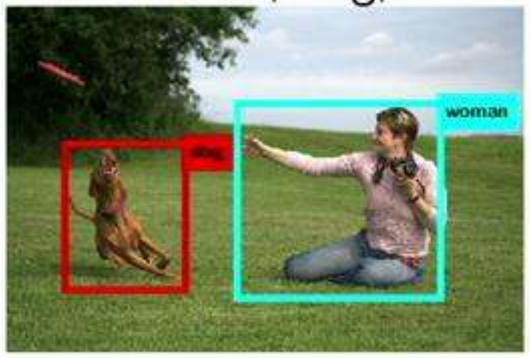
Issues with Existing Visual Captioning Model



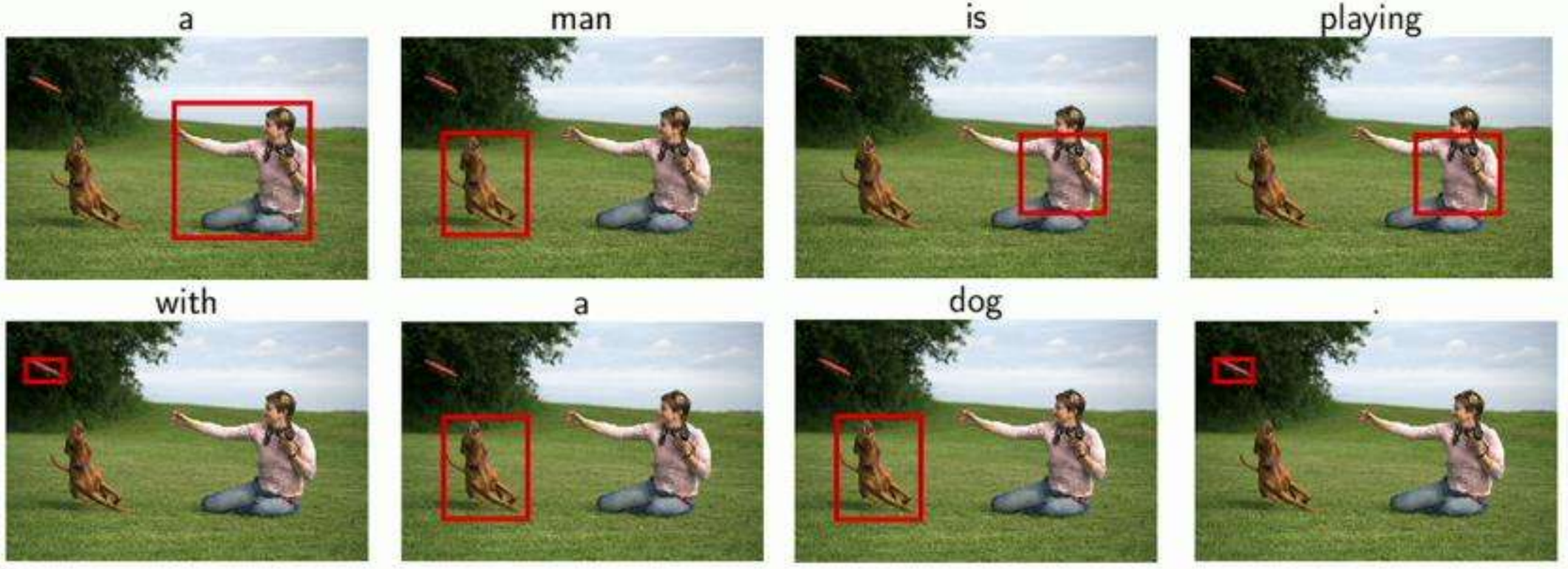
Not properly grounded



Issues with Existing Visual Captioning Model



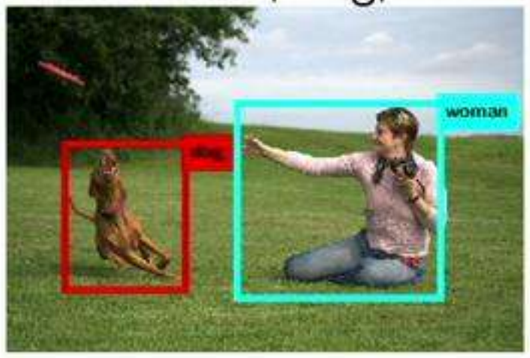
Not properly grounded



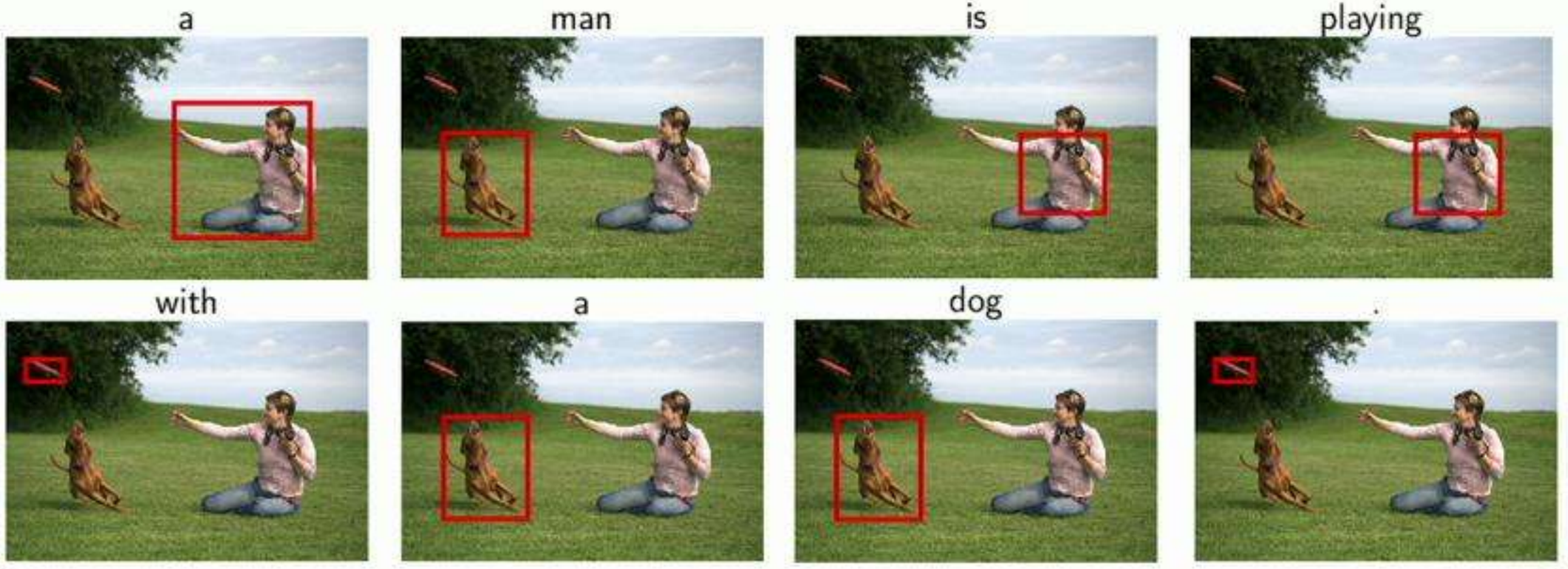
Overly rely on linguistic prior ⟶ Object hallucination

A group of people sitting around a table with laptops.

Issues with Existing Visual Captioning Model



Not properly grounded



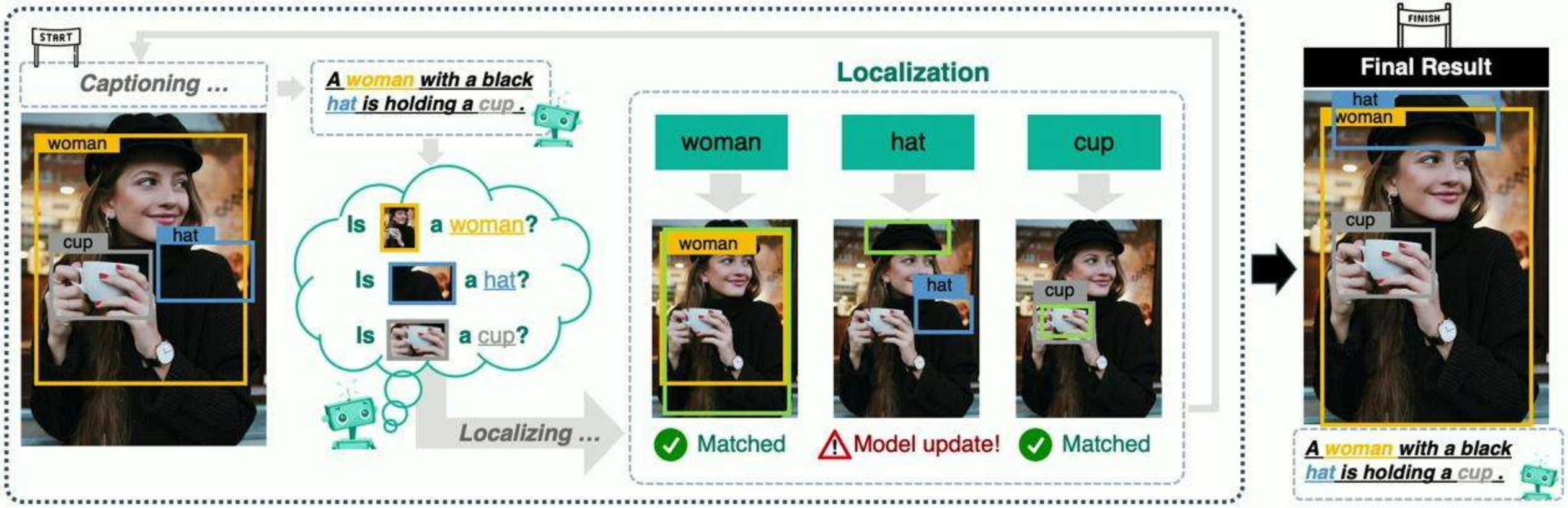
Overly rely on linguistic prior → Object hallucination

A group of people sitting around a table with laptops.

Train with annotations

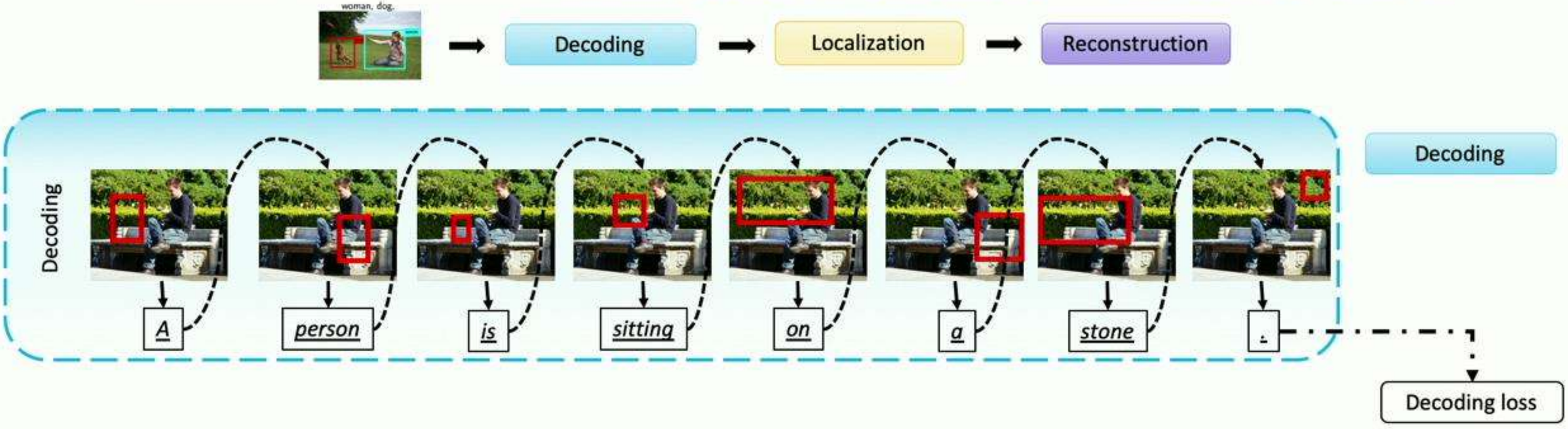
A man with pierced ears is wearing glasses and an orange hat.

Proposed Concept



Cyclical Training Regimen

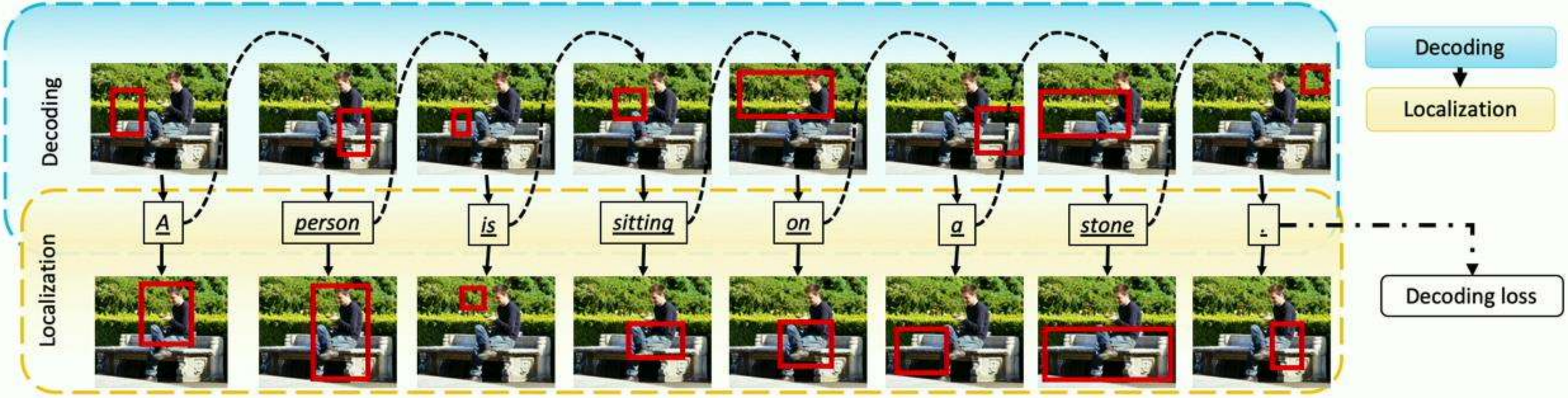
- Grounded Visual Captioning without ground-truth annotations and with a single FC layer.



Cyclical Training Regimen

- Grounded Visual Captioning without ground-truth annotations.

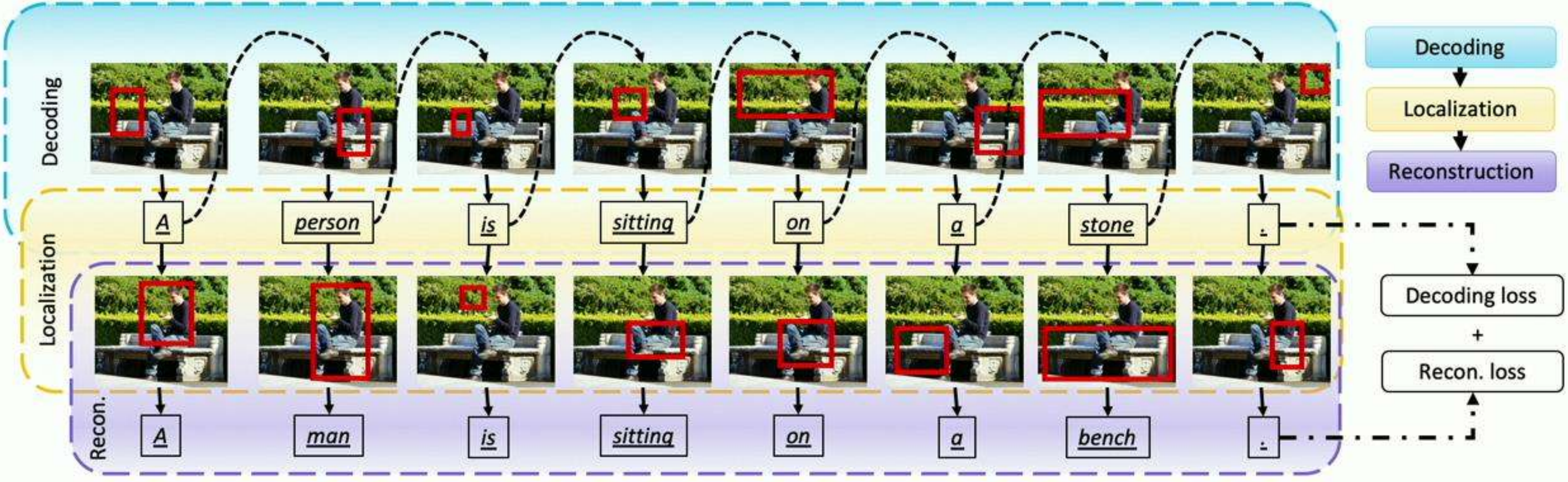
Decoding → Localization → Reconstruction



Cyclical Training Regimen

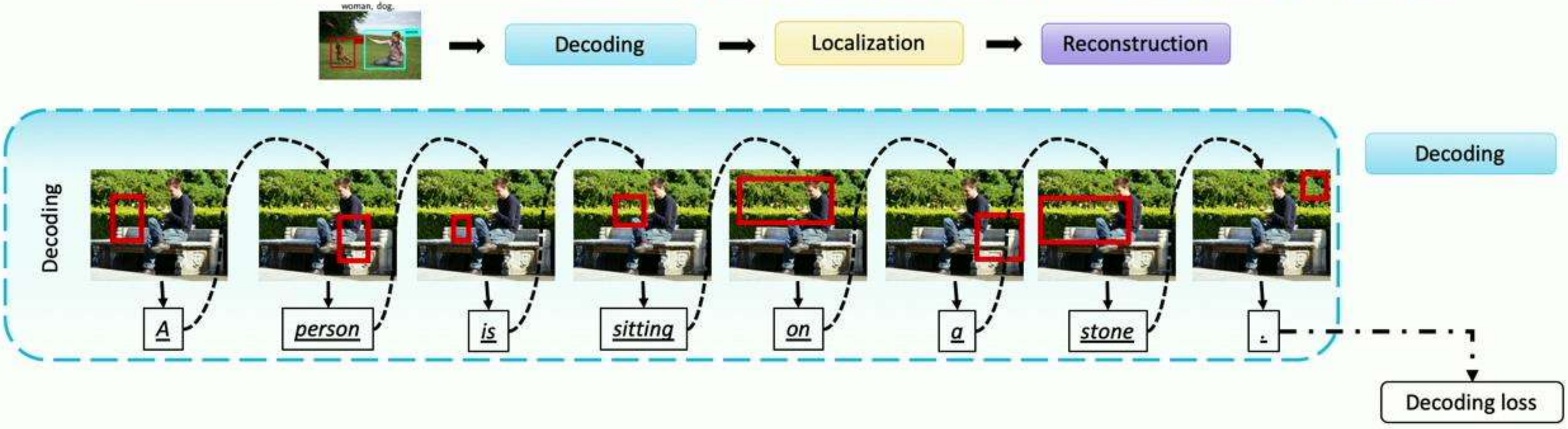
- Grounded Visual Captioning without ground-truth annotations.

Decoding → Localization → Reconstruction

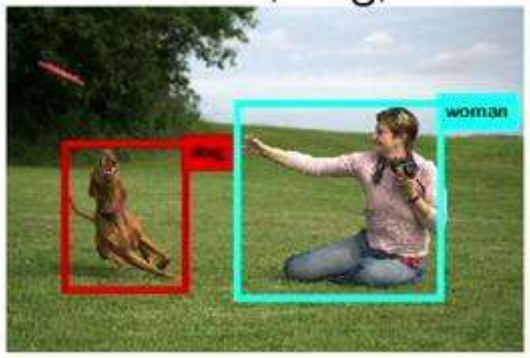


Cyclical Training Regimen

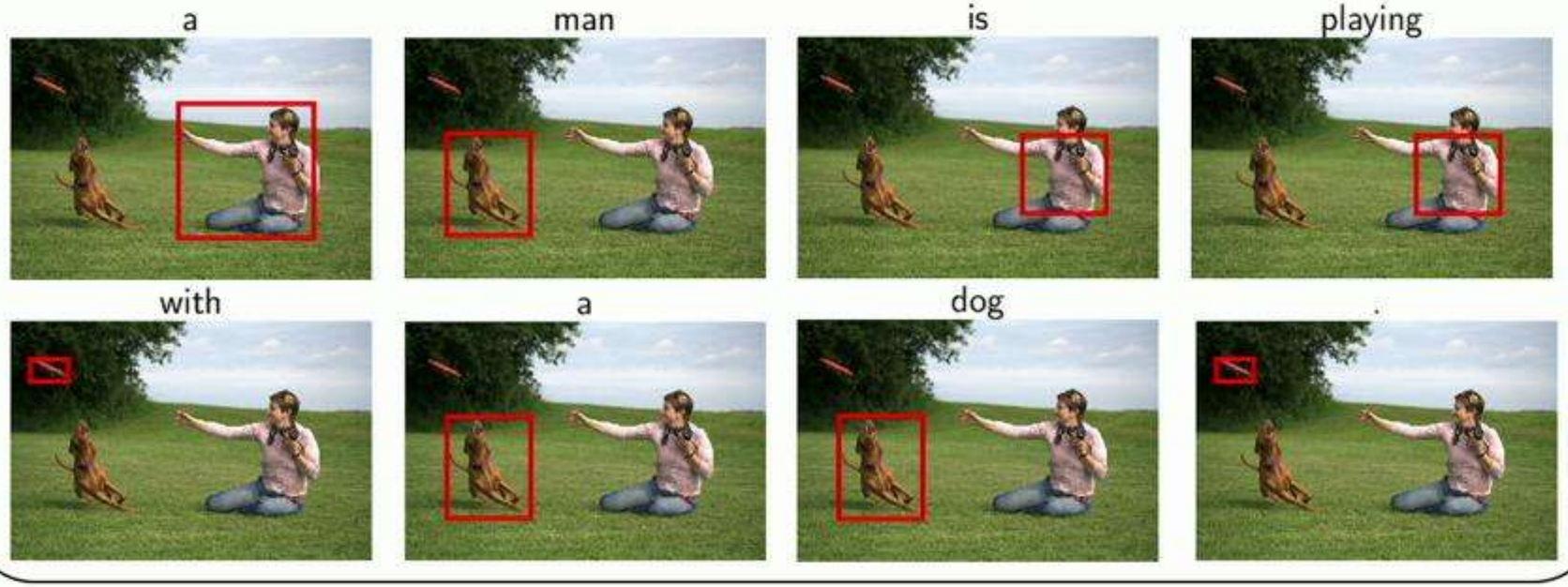
- Grounded Visual Captioning without ground-truth annotations and with a single FC layer.



Issues with Existing Visual Captioning Model



Not properly grounded



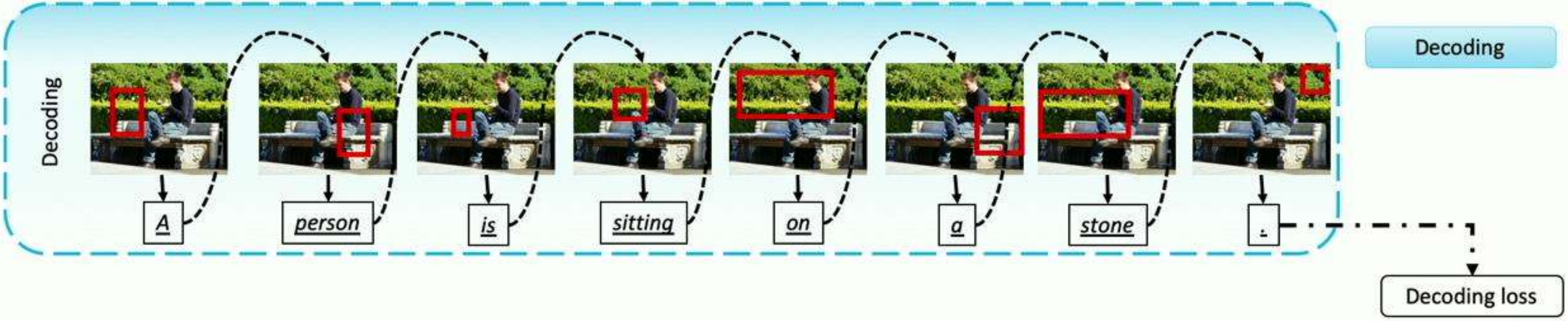
Overly rely on linguistic prior ⟶ Object hallucination

A group of people sitting around a table with laptops.

Cyclical Training Regimen

- Grounded Visual Captioning without ground-truth annotations and with a single FC layer.

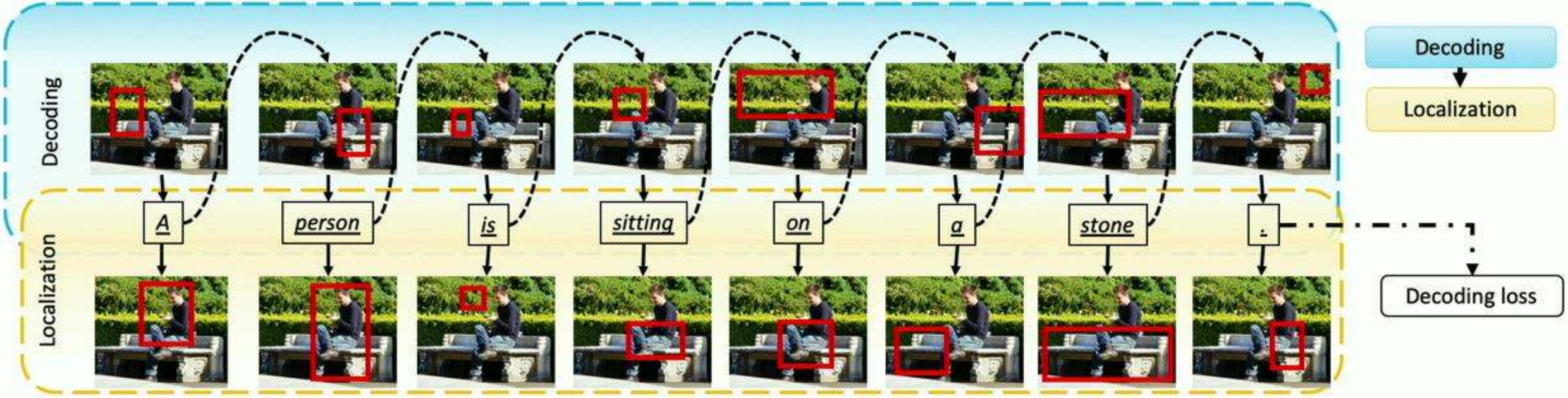
Decoding → Localization → Reconstruction



Cyclical Training Regimen

- Grounded Visual Captioning without ground-truth annotations.

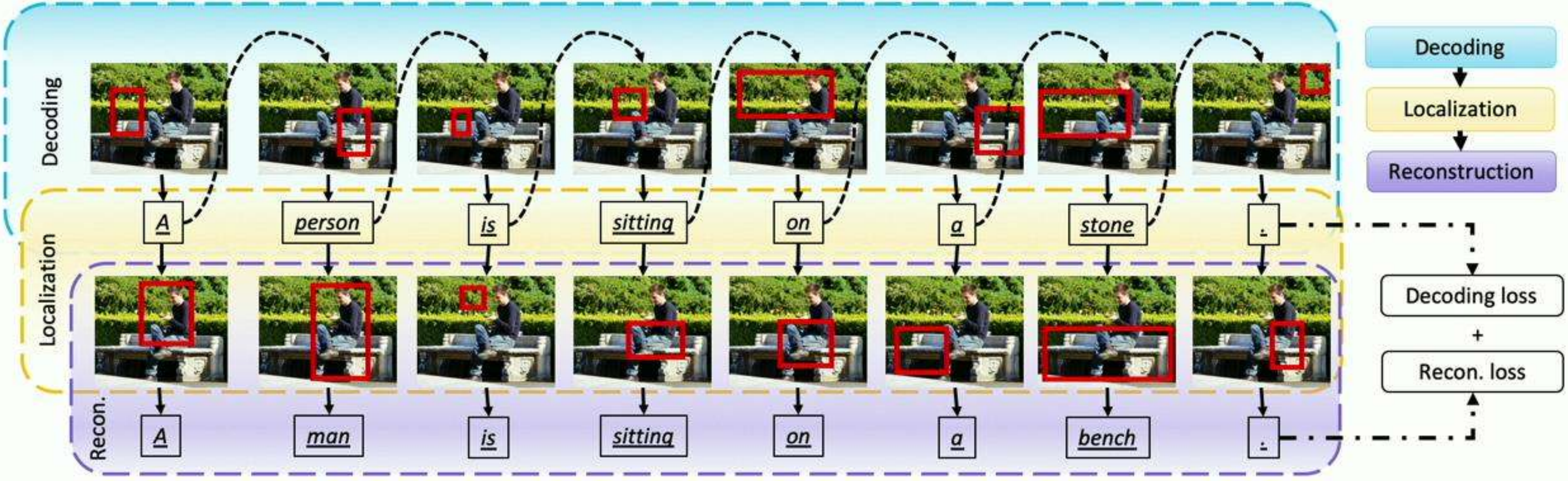
Decoding → Localization → Reconstruction



Cyclical Training Regimen

- Grounded Visual Captioning without ground-truth annotations.

Decoding → Localization → Reconstruction



Experimental Results – Flickr30k Entities

Method		Captioning Evaluation					Grounding Evaluation	
		B@1	B@4	METEOR	CIDEr	SPICE	$F1_{all}$	$F1_{loc}$
ATT-FCN		64.7	19.9	18.5	-	-	-	-
NBT		69.0	27.1	21.7	57.5	15.6	-	-
Up-Down		69.4	27.3	21.7	56.6	16.0	4.14	12.3
GVD	(Sup.)	69.9	27.3	22.5	62.3	16.5	7.77	22.2
	(Unsup.)	69.2	26.9	22.1	60.1	16.1	3.97	11.6
Baseline (avg. 5 runs)	(Sup.)	69.0	26.8	22.4	61.1	16.8	8.44 (+100%)	22.8 (+100%)
	(Unsup.)	69.1	26.0	22.1	59.6	16.3	4.08 (+0%)	11.8 (+0%)
Cyclical (avg. 5 runs)		**69.4**	**26.9**	**22.3**	**60.8**	**16.6**	**5.11 (+24%)**	**14.2 (+22%)**

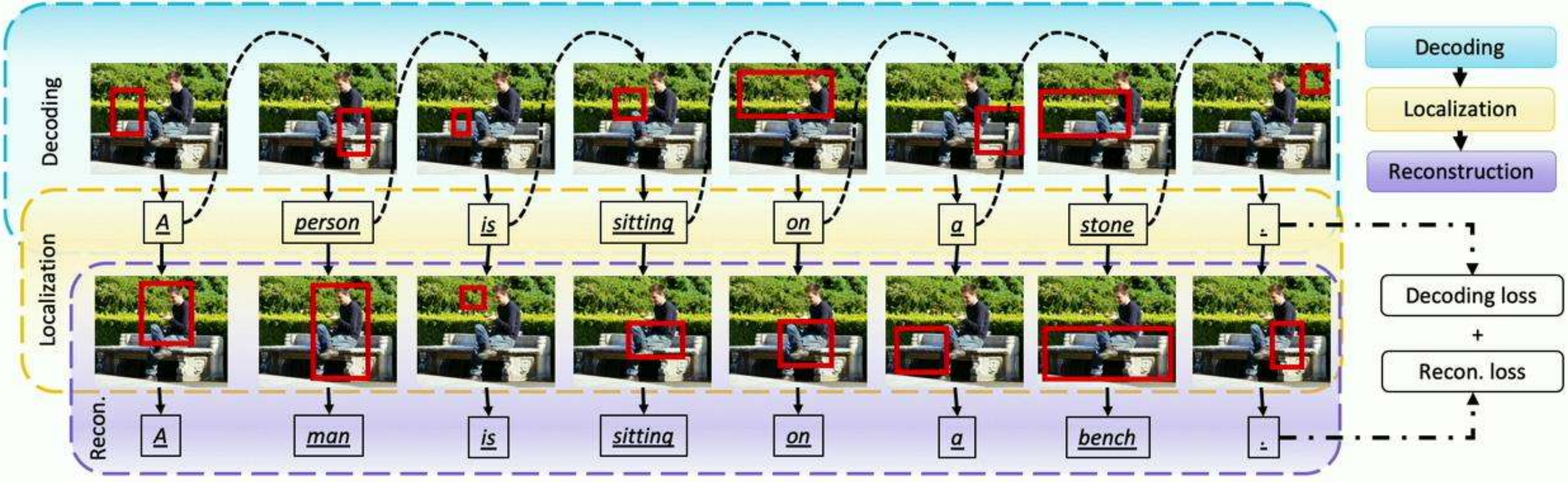
- Only requires to learn one extra FC layer, which can be removed at test time.

ATT-FCN (You et al., CVPR 2016), NBT (Lu et al., CVPR 2018), Up-Down (Anderson et al., CVPR 2018), GVD (Zhou et al., CVPR 2019)

Cyclical Training Regimen

- Grounded Visual Captioning without ground-truth annotations.

Decoding → Localization → Reconstruction



Experimental Results – Flickr30k Entities

Method		Captioning Evaluation					Grounding Evaluation	
		B@1	B@4	METEOR	CIDEr	SPICE	$F1_{all}$	$F1_{loc}$
ATT-FCN		64.7	19.9	18.5	-	-	-	-
NBT		69.0	27.1	21.7	57.5	15.6	-	-
Up-Down		69.4	27.3	21.7	56.6	16.0	4.14	12.3
GVD	(Sup.)	69.9	27.3	22.5	62.3	16.5	7.77	22.2
	(Unsup.)	69.2	26.9	22.1	60.1	16.1	3.97	11.6
Baseline (avg. 5 runs)	(Sup.)	69.0	26.8	22.4	61.1	16.8	8.44 (+100%)	22.8 (+100%)
	(Unsup.)	69.1	26.0	22.1	59.6	16.3	4.08 (+0%)	11.8 (+0%)
Cyclical (avg. 5 runs)		**69.4**	**26.9**	**22.3**	**60.8**	**16.6**	**5.11 (+24%)**	**14.2 (+22%)**

- Only requires to learn one extra FC layer, which can be removed at test time.

ATT-FCN (You et al., CVPR 2016), NBT (Lu et al., CVPR 2018), Up-Down (Anderson et al., CVPR 2018), GVD (Zhou et al., CVPR 2019)

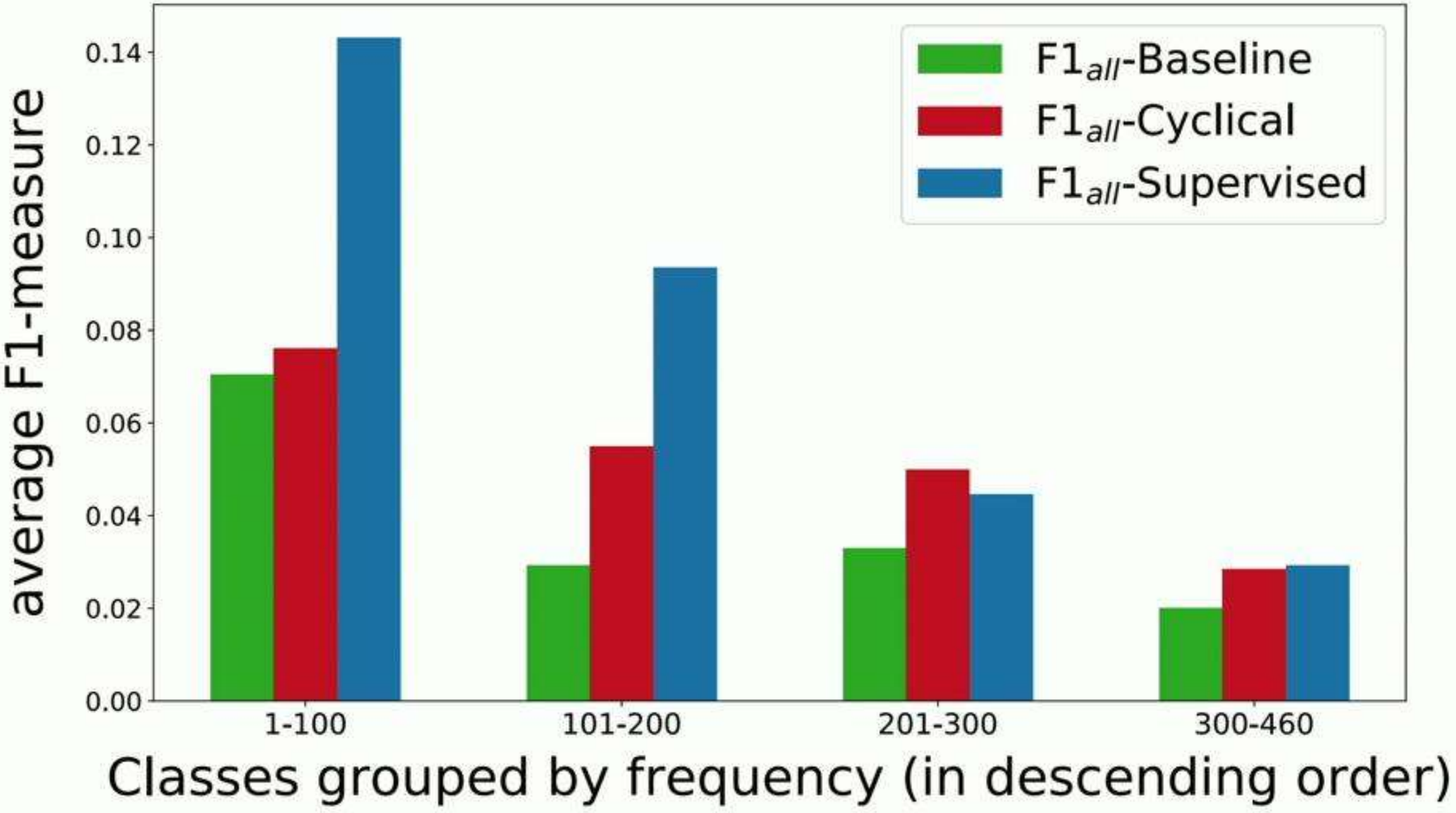
Experimental Results – ActivityNet-Entities

Method		Captioning Evaluation					Grounding Evaluation	
		B@1	B@4	METEOR	CIDEr	SPICE	$F1_{all}$	$F1_{loc}$
GVD	(Sup.)	23.9	2.59	11.2	47.5	15.1	7.11	24.1
	(Unsup.) w/o SelfAttn	23.2	2.28	10.9	45.6	15.0	3.70	12.7
Baseline (avg. 5 runs)	(Sup.)	23.1	2.13	10.7	45.0	14.6	7.30 (+100%)	25.0 (+100%)
	(Unsup.)	23.2	2.22	10.8	45.9	**15.1**	3.75 (+0%)	12.0 (+0%)
Cyclical (avg. 5 runs)		**23.7**	**2.45**	**11.1**	**46.4**	14.8	**4.68 (+26%)**	**15.8 (+22%)**

- Only requires to learn one extra FC layer, which can be removed at test time.

GVD (Zhou et al., CVPR 2019)

Grounding Acc. VS Annotation Frequency



Qualitative Results

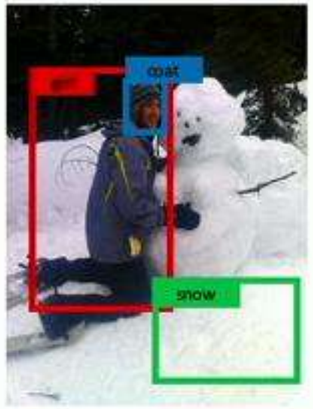
Baseline: *A group of people are watching a game.*

Proposed: *A group of men in white uniforms are standing in a field with a crowd watching.*

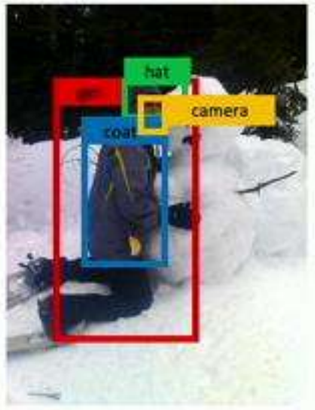
Qualitative Results

Baseline: *A group of people are watching a game.*

Proposed: *A group of men in white uniforms are standing in a field with a crowd watching.*

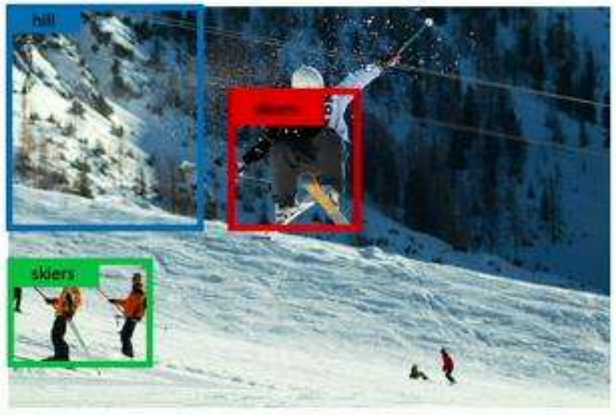


Baseline: *A young girl in a blue coat is sitting in the snow.*



Proposed: *A young girl wearing a winter hat and a purple coat is smiling at the camera.*

Baseline: *Four skiers are skiing down a snowy mountain.*

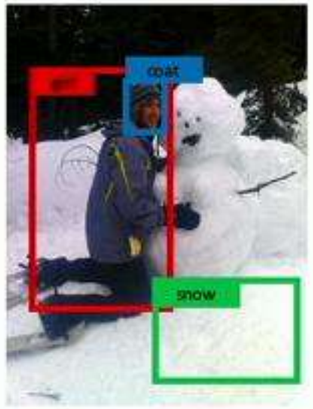


Proposed: *A skier is jumping over a snowy hill while other skiers watch.*

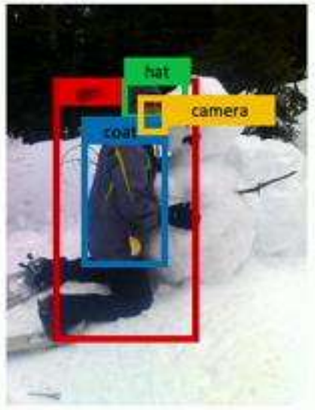
Qualitative Results

Baseline: *A group of people are watching a game.*

Proposed: *A group of men in white uniforms are standing in a field with a crowd watching.*

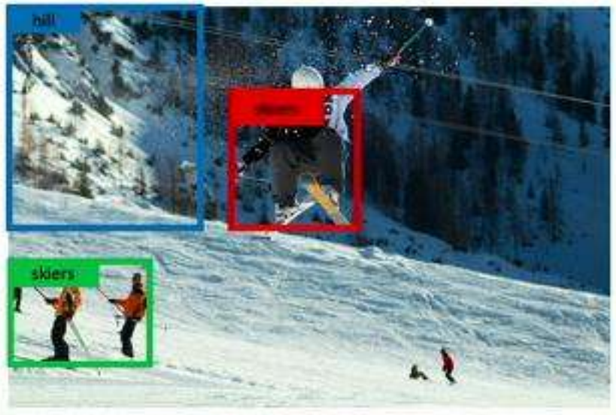


Baseline: *A young girl in a blue coat is sitting in the snow.*



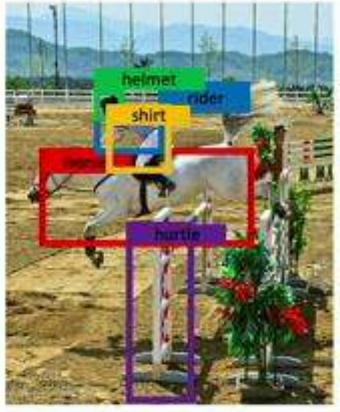
Proposed: *A young girl wearing a winter hat and a purple coat is smiling at the camera.*

Baseline: *Four skiers are skiing down a snowy mountain.*



Proposed: *A skier is jumping over a snowy hill while other skiers watch.*

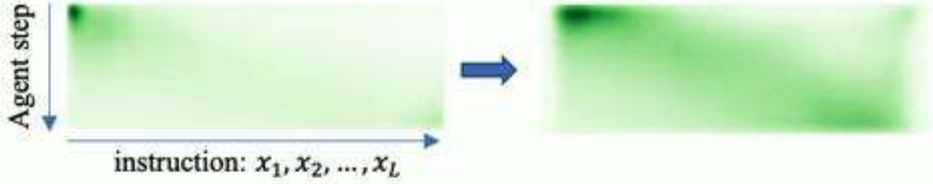
Baseline: *A white horse is jumping over an obstacle.*



Proposed: *A white horse with a rider in a blue helmet and white shirt jumping over a hurtle.*

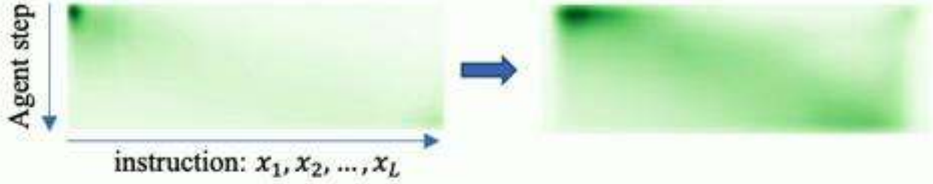
Summary of Contributions

- A self-monitoring mechanism for grounding on the instructions
  - Introduce end-to-end learned agent for backtracking

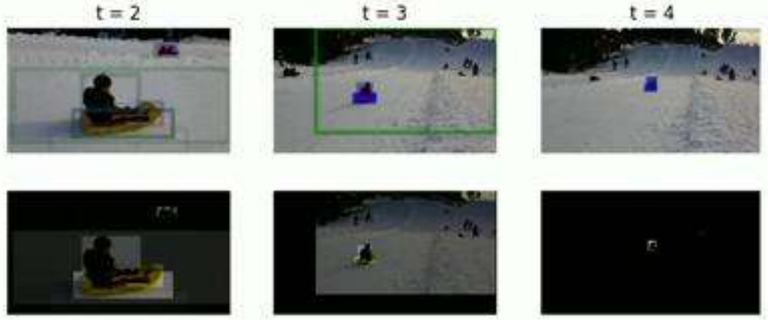


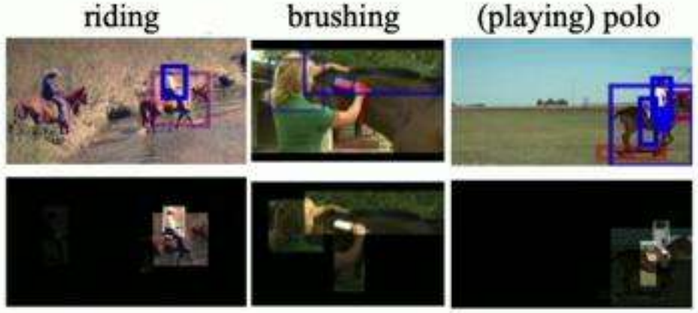
Summary of Contributions

- A self-monitoring mechanism for grounding on the instructions
  - Introduce end-to-end learned agent for backtracking

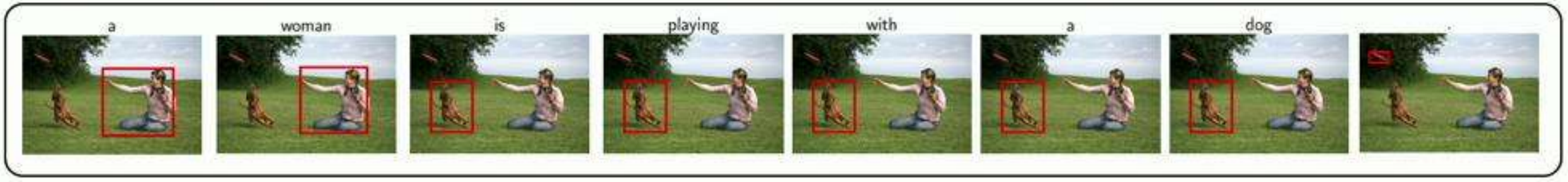


- Object-level understanding for both action recognition and video captioning
  - Ground video understanding tasks to objects and interactions





- Propose a cyclical training regimen for grounded visual captioning



Publications

- **Chih-Yao Ma**, Yannis Kalantidis, Ghassan AlRegib, Peter Vajda, Marcus Rohrbach, Zsolt Kira
  Learning to Generate Grounded Visual Captions without Localization Supervision
  Technical Report, 2019
- Chia-Wen Kuo, **Chih-Yao Ma**, Jia-Bin Huang, Zsolt Kira
  Manifold Graph with Learned Prototypes for Semi-Supervised Image Classification
  Technical Report, 2019
- Yen-Cheng Liu, Junjiao Tian, **Chih-Yao Ma**, Chia-Wen Kuo, Zsolt Kira
  Bandwidth-limited Collaborative Agents for Perception
  Technical Report, 2019
- **Chih-Yao Ma**, Zuxuan Wu, Ghassan AlRegib, Caiming Xiong, Zsolt Kira
  The Regretful Agent: Heuristic-Aided Navigation through Progress Estimation
  Computer Vision and Pattern Recognition (CVPR), 2019 **(Oral)**
- Zuxuan Wu, Caiming Xiong, **Chih-Yao Ma**, Richard Socher, Larry S Davis
  AdaFrame: Adaptive Frame Selection for Fast Video Recognition
  Computer Vision and Pattern Recognition (CVPR), 2019
- **Chih-Yao Ma**, Jiasen Lu, Zuxuan Wu, Ghassan AlRegib, Zsolt Kira, Richard Socher, Caiming Xiong
  Self-Monitoring Navigation Agent via Auxiliary Progress Estimation
  International Conference on Learning Representations (ICLR), 2019 ***(Top 7% of reviews)***
- **Chih-Yao Ma**, Asim Kadav, Iain Melvin, Zsolt Kira, Ghassan AlRegib, Hans Peter Graf
  Attend and Interact: Higher-Order Object Interactions for Video Understanding
  Computer Vision and Pattern Recognition (CVPR), 2018
- **Chih-Yao Ma***, Min-Hung Chen*, Zsolt Kira, and Ghassan AlRegib
  TS-LSTM and temporal-inception: Exploiting spatiotemporal dynamics for activity recognition
  Signal Processing: Image Communication, 2018 (* equal contribution)
- **Chih-Yao Ma**, Asim Kadav, Iain Melvin, Zsolt Kira, Ghassan AlRegib, Hans Peter Graf
  Grounded Objects and Interactions for Video Captioning
  Neural Information Processing Systems (NeurIPS) Workshop on Visually-Grounded Interaction and Language, 2017

Collaborators

Georgia Tech

salesforce research

NEC

facebook research

Ghassan AlRegib

Zsolt Kira

Caiming Xiong

Richard Socher

Hans Peter Graf

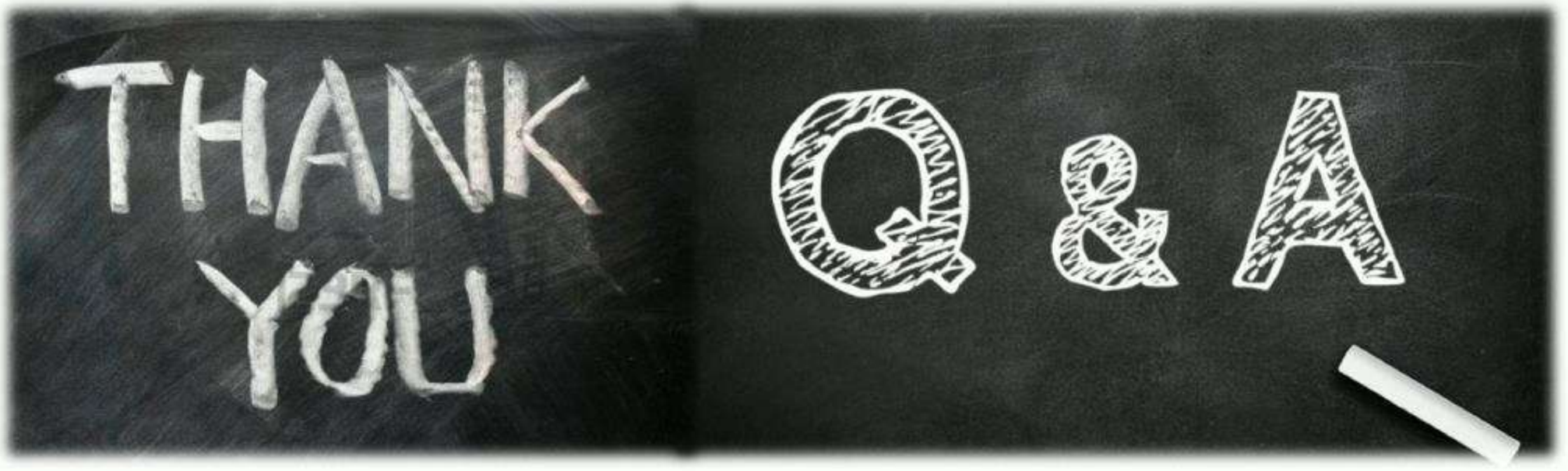
Asim Kadav

Iain Melvin

Marcus Rohrbach

Yannis Kalantidis

Peter Vajda


THANK
YOU
Q & A