The Situated Interactive Language Grounding Benchmark

Victor Zhong, Austin W. Hanjie, Sida I. Wang, Karthik Narasimhan, Luke Zettlemoyer

UWNLP

PRINCETON NLP

FACEBOOK AI

NeurIPS 2021

Motivation

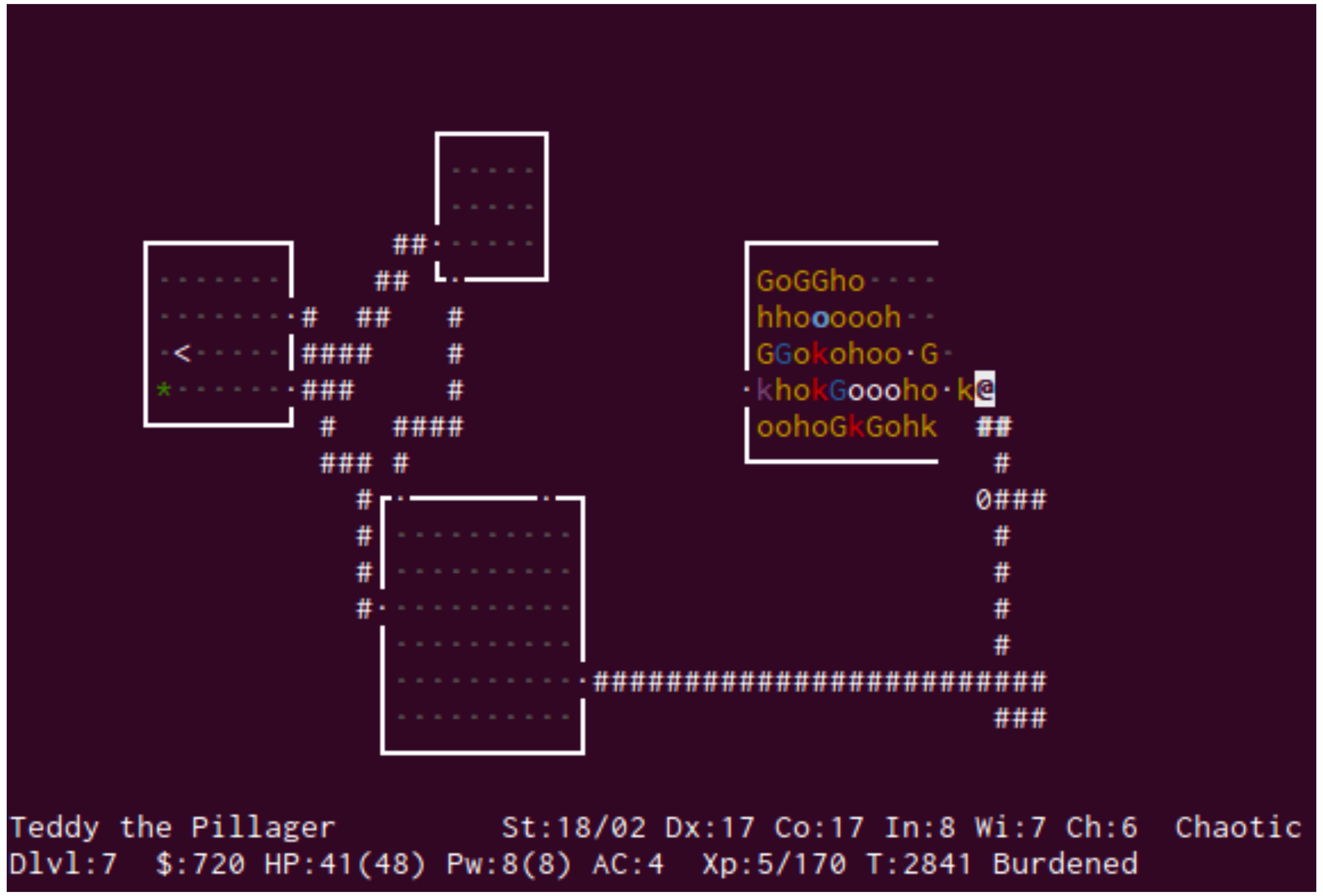
What do we want in “language grounding”?

- Goal: reading agents should generalize across different phenomena

Motivation

What do we want in “language grounding”?

- Goal: reading agents should generalize across different phenomena
  - Complex scenes (rich visual, sophisticated procgen, partial observability)

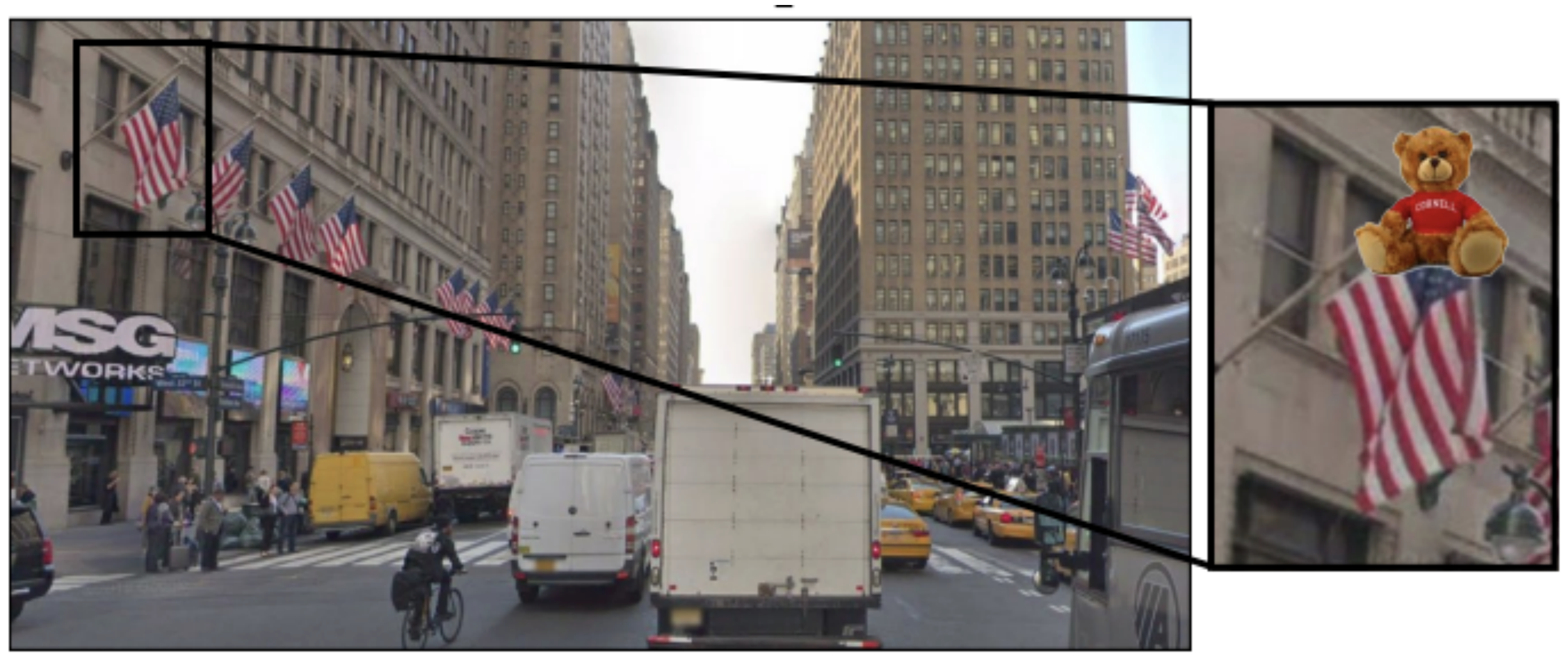


```
Teddy the Pillager          St:18/02 Dx:17 Co:17 In:8 Wi:7 Ch:6  Chaotic
Dlvl:7  $:720 HP:41(48) Pw:8(8) AC:4  Xp:5/170 T:2841 Burdened
```

Motivation

What do we want in “language grounding”?

- Goal: reading agents should generalize across different phenomena
    - Complex scenes (rich visual, sophisticated procgen, partial observability)
    - New natural language references

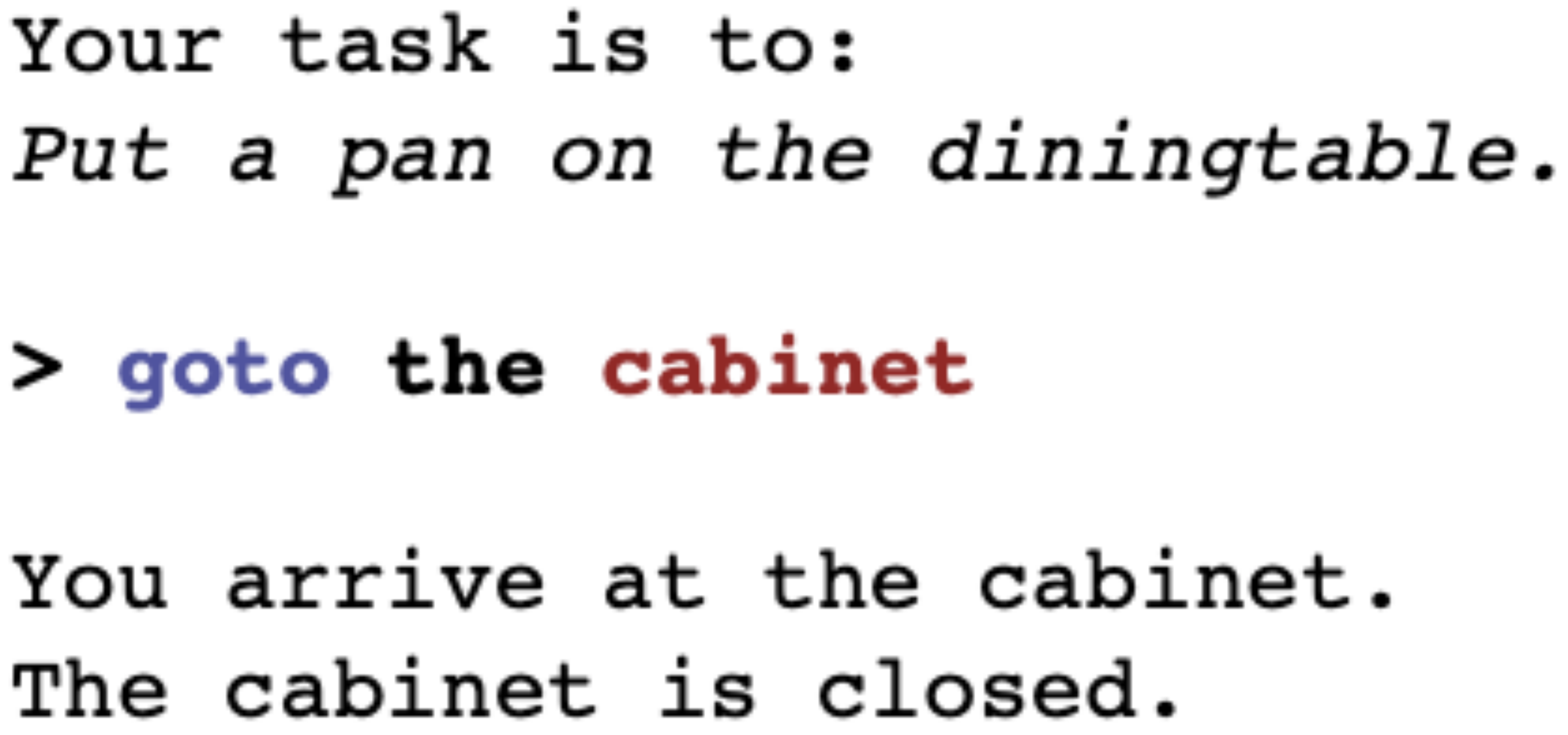


Turn and go with the flow of traffic. At the first traffic light turn left. Go past the next two traffic light, As you come to the third traffic light you will see a white building on your left with many American flags on it. Touchdown is sitting in the stars of the first flag.

Motivation

What do we want in “language grounding”?

- Goal: reading agents should generalize across different phenomena
    - Complex scenes (rich visual, sophisticated procgen, partial observability)
    - New natural language references
    - Large language action space

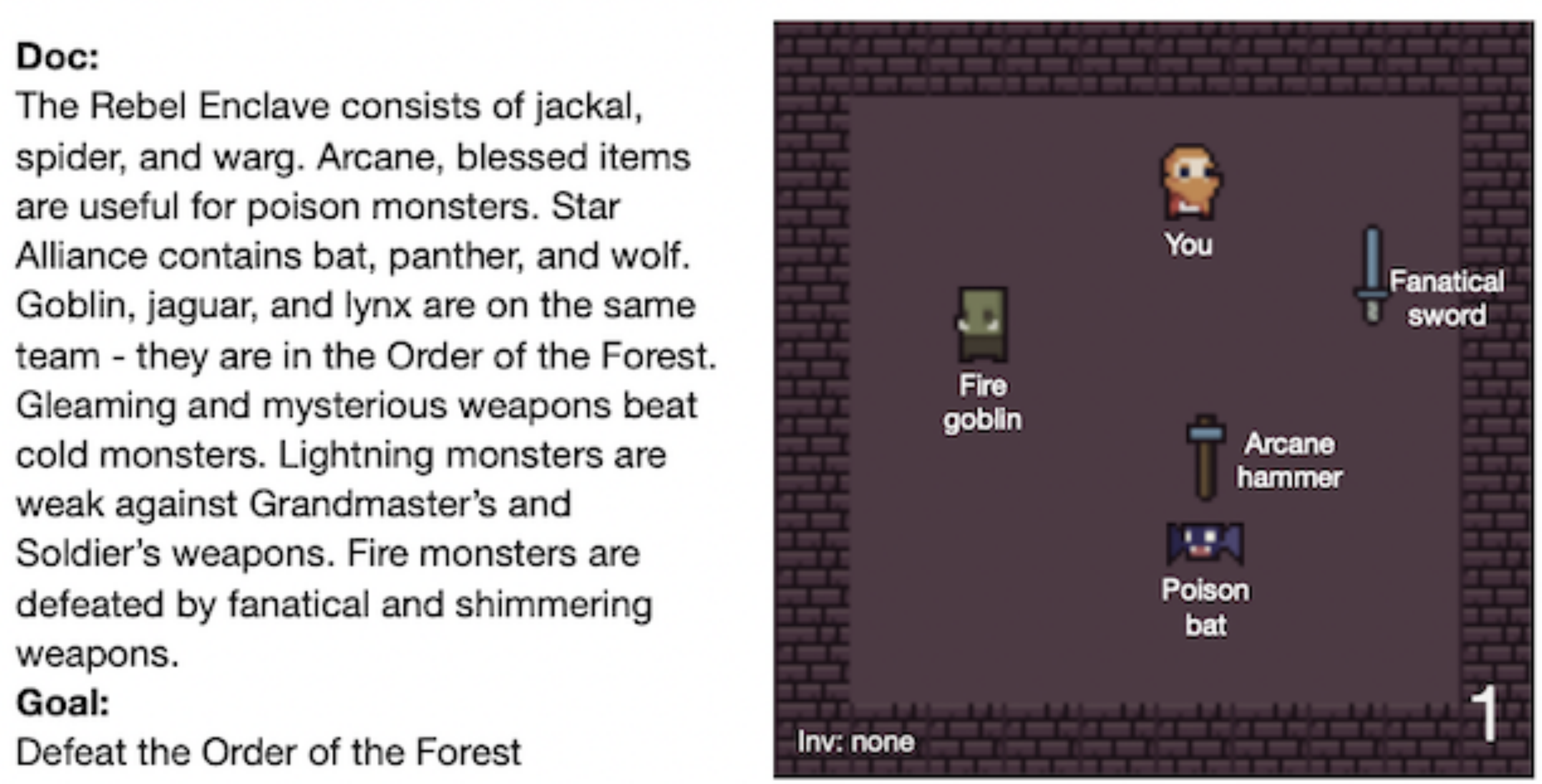


```
Your task is to:
Put a pan on the diningtable.

> goto the cabinet

You arrive at the cabinet.
The cabinet is closed.
```

Motivation

What do we want in “language grounding”?

- Goal: reading agents should generalize across different phenomena
  - Complex scenes (rich visual, sophisticated procgen, partial observability)
  - New natural language references
  - Large language action space
  - Multi-hop references/reasoning

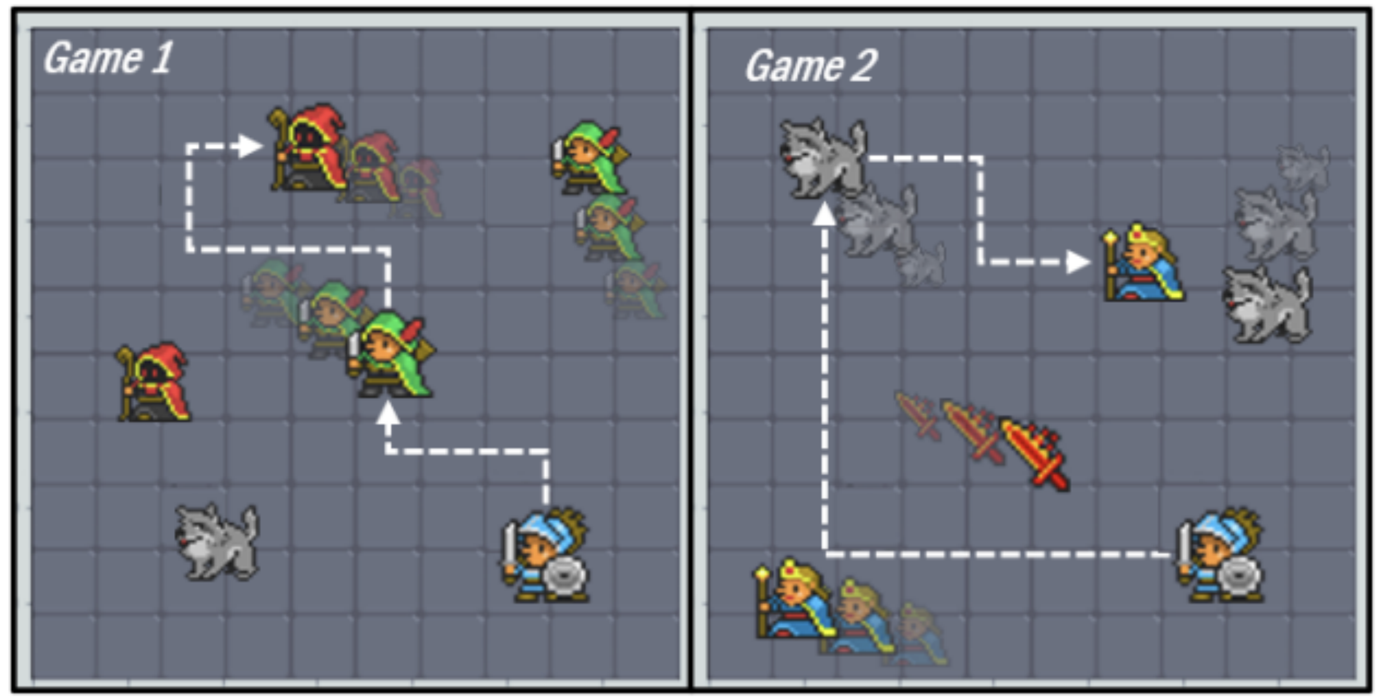


Doc:
The Rebel Enclave consists of jackal, spider, and warg. Arcane, blessed items are useful for poison monsters. Star Alliance contains bat, panther, and wolf. Goblin, jaguar, and lynx are on the same team - they are in the Order of the Forest. Gleaming and mysterious weapons beat cold monsters. Lightning monsters are weak against Grandmaster’s and Soldier’s weapons. Fire monsters are defeated by fanatical and shimmering weapons.
Goal:
Defeat the Order of the Forest

Motivation

What do we want in “language grounding”?

- Goal: reading agents should generalize across different phenomena
  - Complex scenes (rich visual, sophisticated procgen, partial observability)
  - New natural language references
  - Large language action space
  - Multi-hop references/reasoning
  - New entity dynamics



GAME 1 MANUAL

1. at a particular locale, there exists a motionless mongrel that is a formidable adversary.
2. the top-secret paperwork is in the crook’s possession, and he’s heading closer and closer to where you are.
3. the crucial target is held by the wizard and the wizard is fleeing from you.
4. the mugger rushing away is the opposition posing a serious threat.
5. the thing that is not able to move is the mage who possesses the enemy that is deadly.
6. *the vital goal is found with the canine, but it is running away from you.*

Motivation

What do we want in “language grounding”?

- Goal: reading agents should generalize across different phenomena
  - Complex scenes (rich visual, sophisticated procgen, partial observability)
  - New natural language references
  - Large language action space
  - Multi-hop references/reasoning
  - New entity dynamics
- Currently: test on single environment that emphasize particular phenomena

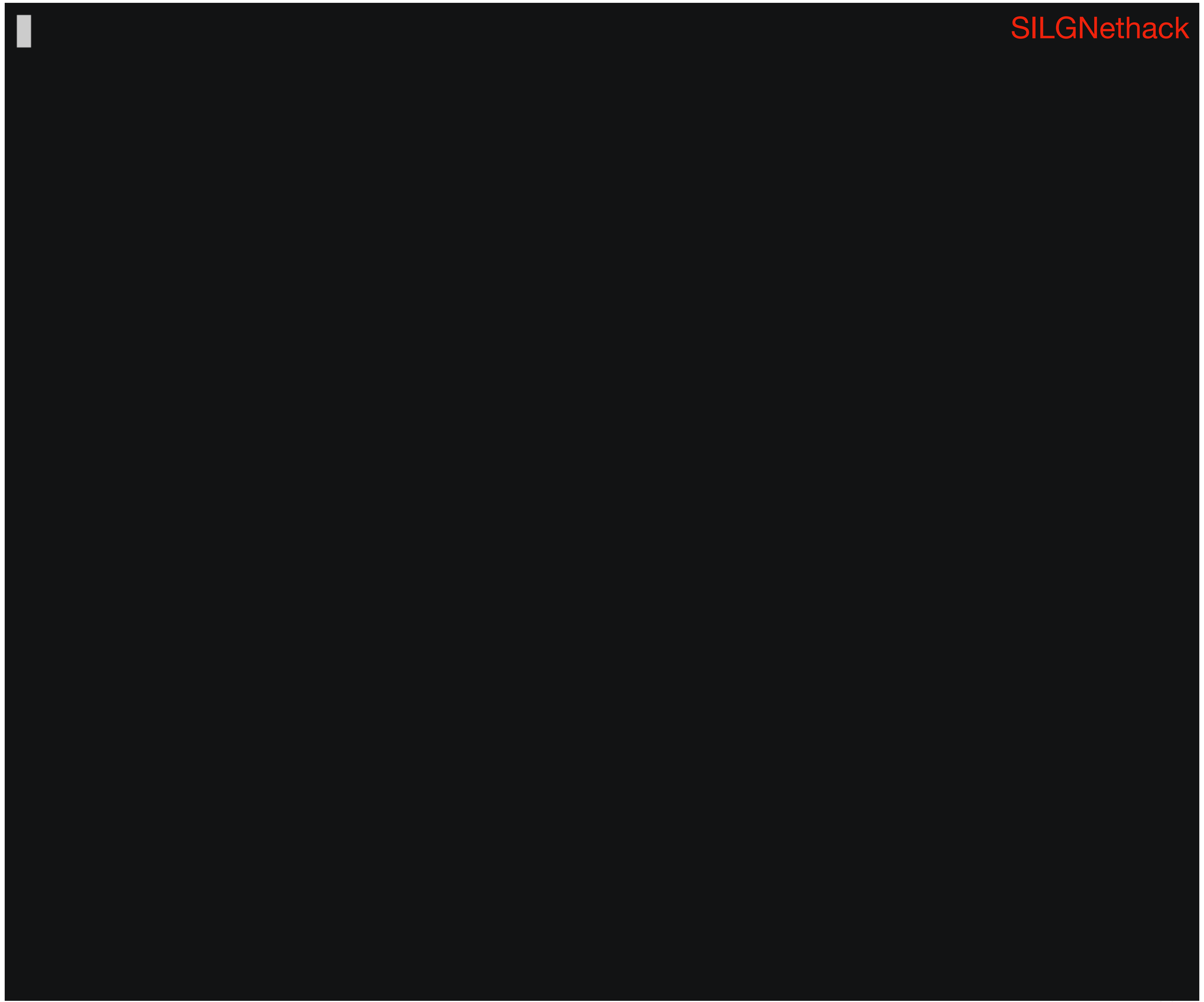
Proposal

SILG: Situated Interactive Language Grounding Benchmark

- Combines unique generalization challenges
    - Complex scenes (SILGNethack, ALFWorld, SILGTouchdown)
    - New natural language references (Messenger, SILGTouchdown)
    - Large language action space (ALFWorld)
    - Multi-hop references/reasoning (RTFM, SILGTouchdown)
    - New entity dynamics (RTFM, Messenger)

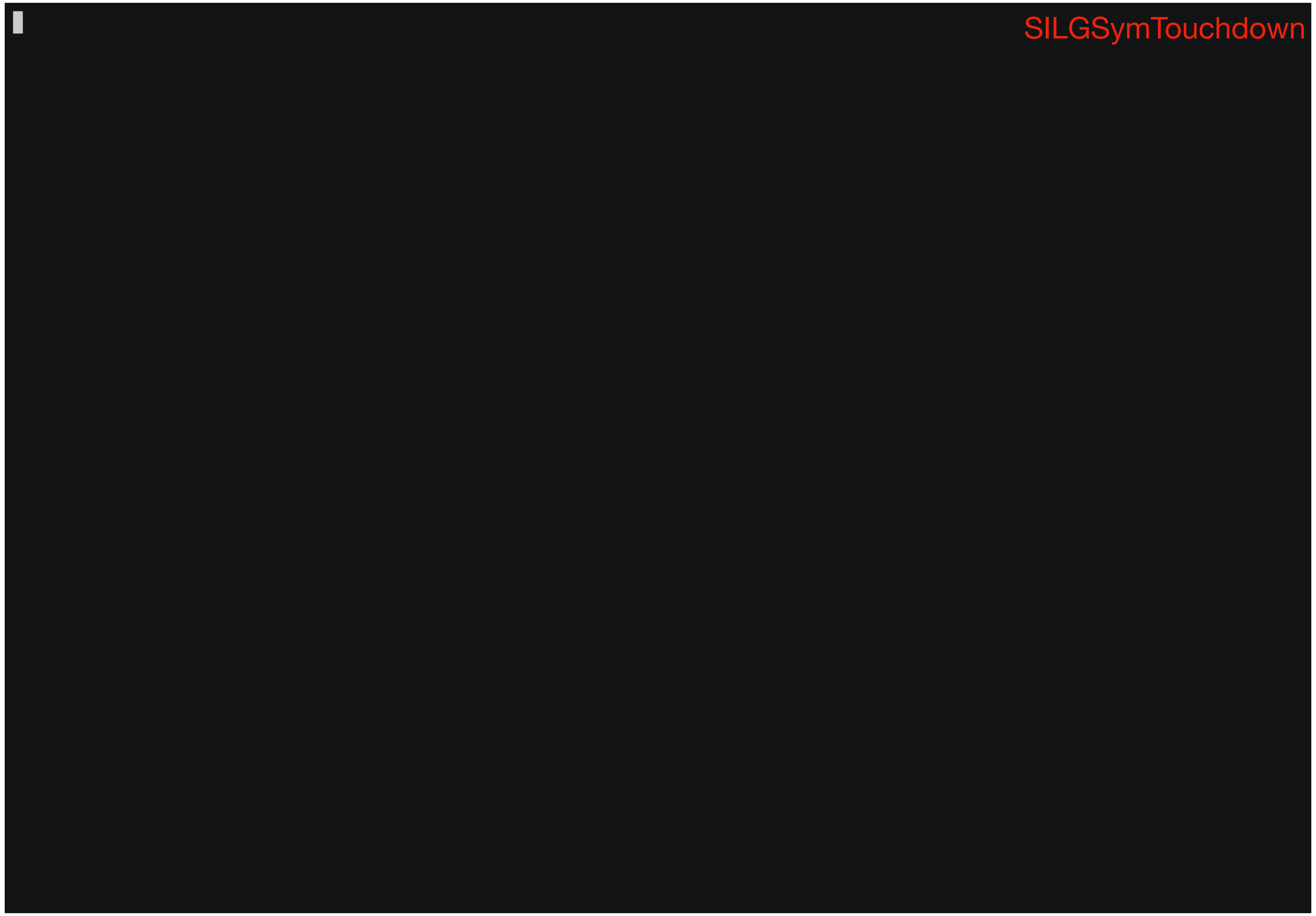
SILG Environment demos

You can play these with yourself!



SILG Environment demos

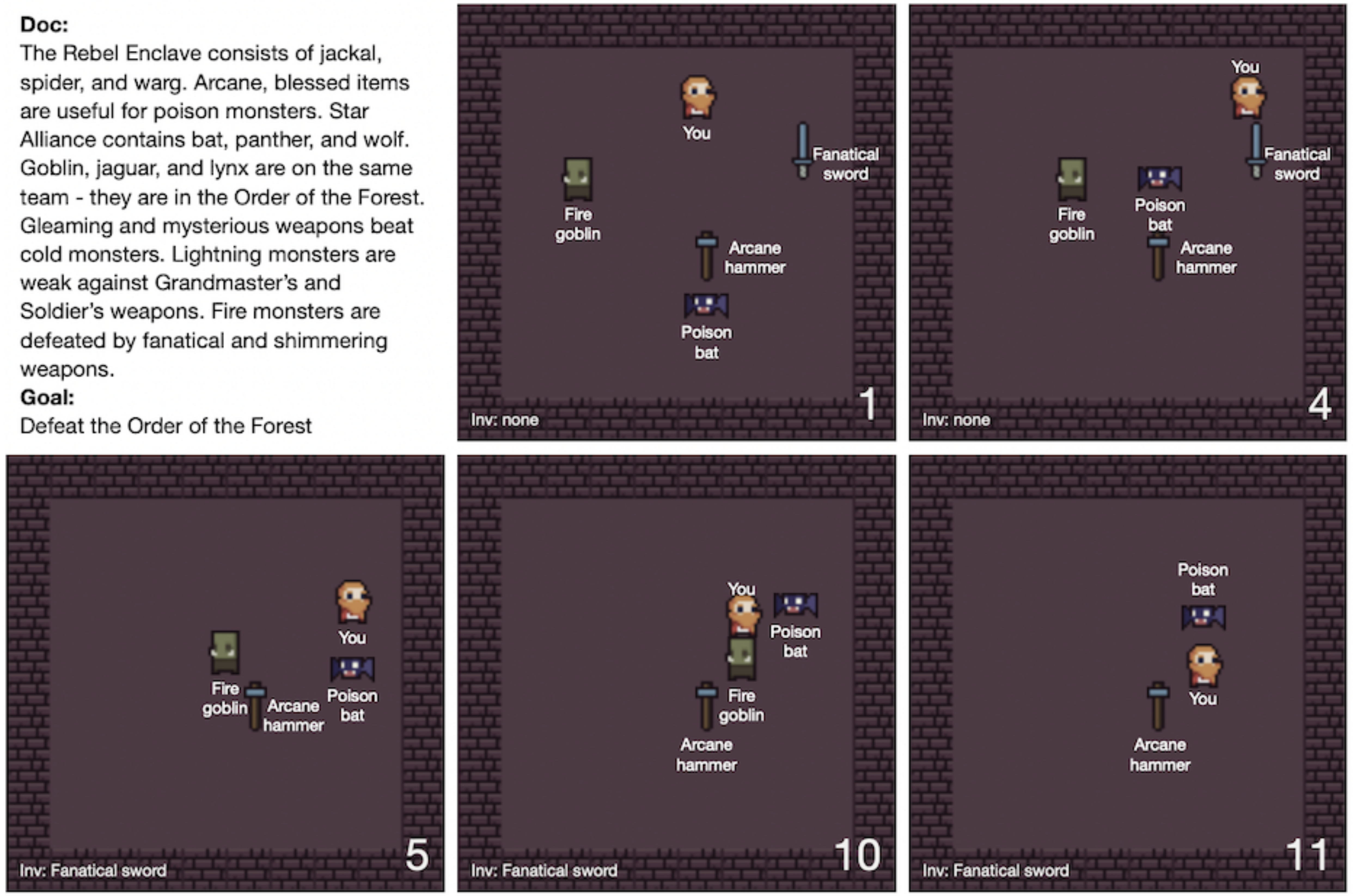
You can play these with yourself!



Implementation

Included environments: RTFM

- Read text describing high level rules to obtain correct item and fight correct monster
- Challenges
  - Combine multi-hop references b/w text and world observations to generalize to worlds with new rules



Doc:
The Rebel Enclave consists of jackal, spider, and warg. Arcane, blessed items are useful for poison monsters. Star Alliance contains bat, panther, and wolf. Goblin, jaguar, and lynx are on the same team - they are in the Order of the Forest. Gleaming and mysterious weapons beat cold monsters. Lightning monsters are weak against Grandmaster's and Soldier's weapons. Fire monsters are defeated by fanatical and shimmering weapons.

Goal:
Defeat the Order of the Forest

RTFM: Generalising to Novel Environment Dynamics via Reading
Victor Zhong, Tim Rocktäschel, and Edward Grefenstette
ICLR 2020

Implementation

Included environments: RTFM

- Read text describing high level rules to obtain correct item and fight correct monster
- Challenges
  - Combine multi-hop references b/w text and world observations to generalize to worlds with new rules

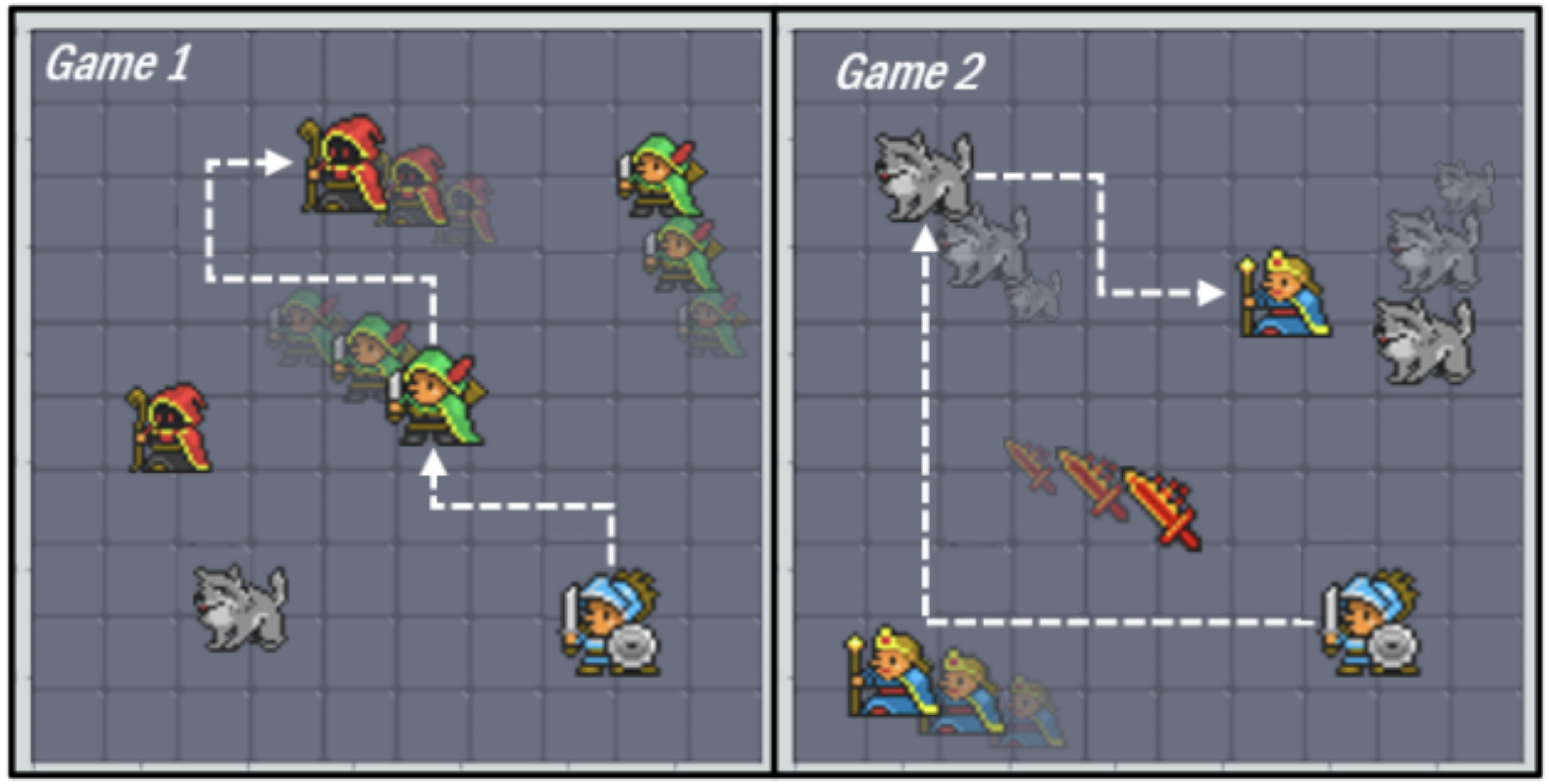
- Included curriculum stages
  - S1: stationary, no distractors, fixed language
  - S2: S1 + movement
  - S3: S2 + distractors
  - S4: S3 + random language templates

RTFM: Generalising to Novel Environment Dynamics via Reading
Victor Zhong, Tim Rocktäschel, and Edward Grefenstette
ICLR 2020

Implementation

Included environments: Messenger

- Read text describing high level rules to visit entities in correct order
- Challenges
  - Ground NL references to entity IDs in scene
  - Generalization given small training distribution of envs



GAME 1 MANUAL

1. at a particular locale, there exists a motionless mongrel that is a formidable adversary.
2. the top-secret paperwork is in the crook's possession, and he's heading closer and closer to where you are.
3. the crucial target is held by the wizard and the wizard is fleeing from you.
4. the mugger rushing away is the opposition posing a serious threat.
5. the thing that is not able to move is the mage who possesses the enemy that is deadly.
6. *the vital goal is found with the canine, but it is running away from you.*

Grounding Language to Entities and Dynamics for Generalization in Reinforcement Learning
Austin Wang, Victor Zhong, Karthik Narasimhan
ICML 2021

Implementation

Included environments: Messenger

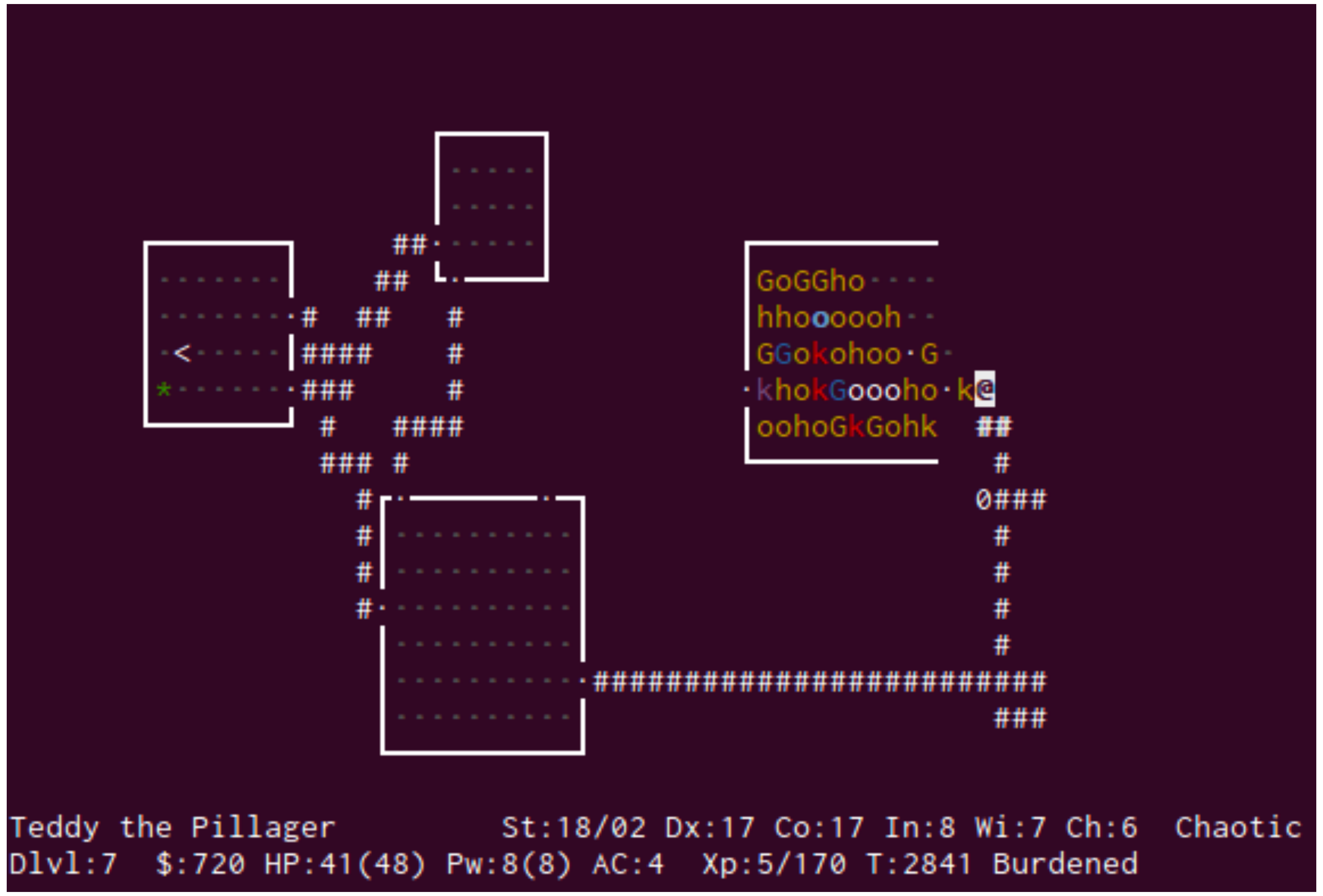
- Read text describing high level rules to visit entities in correct order
- Challenges
  - Ground NL references to entity IDs in scene
  - Generalization given small training distribution of envs

- Included curriculum stages
  - S1: 3 entities, stationary, 1 reasoning step
  - S2: S1 + movement, 1-2 steps
  - S3: S2 + 5 entities, distractors

Implementation

Included environments: SILGNethack

- Read goal specification and env feedback to perform long-horizon task in large procgen world
- Challenges
    - Partially observed, complex world
    - Very delayed, sparse reward



```
                        ┌─────┐
                        │.....│
                        │.....│
       ┌───────┐      ##......│         ┌───────────
       │.......│    ##  └.────┘         │GoGGho....
       │........#  ##   #               │hhoooooh..
       │.<.....│####    #               │GGokohoo.G.
       │*.......###     #               .khokGoooho.k@
       └───────┘ #   ####               │oohoGkGohk  ##
                 ### #                  └───────────  #
                   #┌.───────.─┐                     0###
                   #│..........│                      #
                   #│..........│                      #
                   #...........│                      #
                    │..........│                      #
                    │...........#########################
                    │..........│                      ###
                    └──────────┘
Teddy the Pillager         St:18/02 Dx:17 Co:17 In:8 Wi:7 Ch:6  Chaotic
Dlvl:7  $:720 HP:41(48) Pw:8(8) AC:4  Xp:5/170 T:2841 Burdened
```

The Nethack Learning Environment
Heinrich Küttler, Nantas Nardelli, Alexander H. Miller, Roberta Raileanu, Marco Selvatici, Edward Grefenstette, Tim Rocktäschel
NeurIPS 2020

Implementation

Included environments: SILGNethack

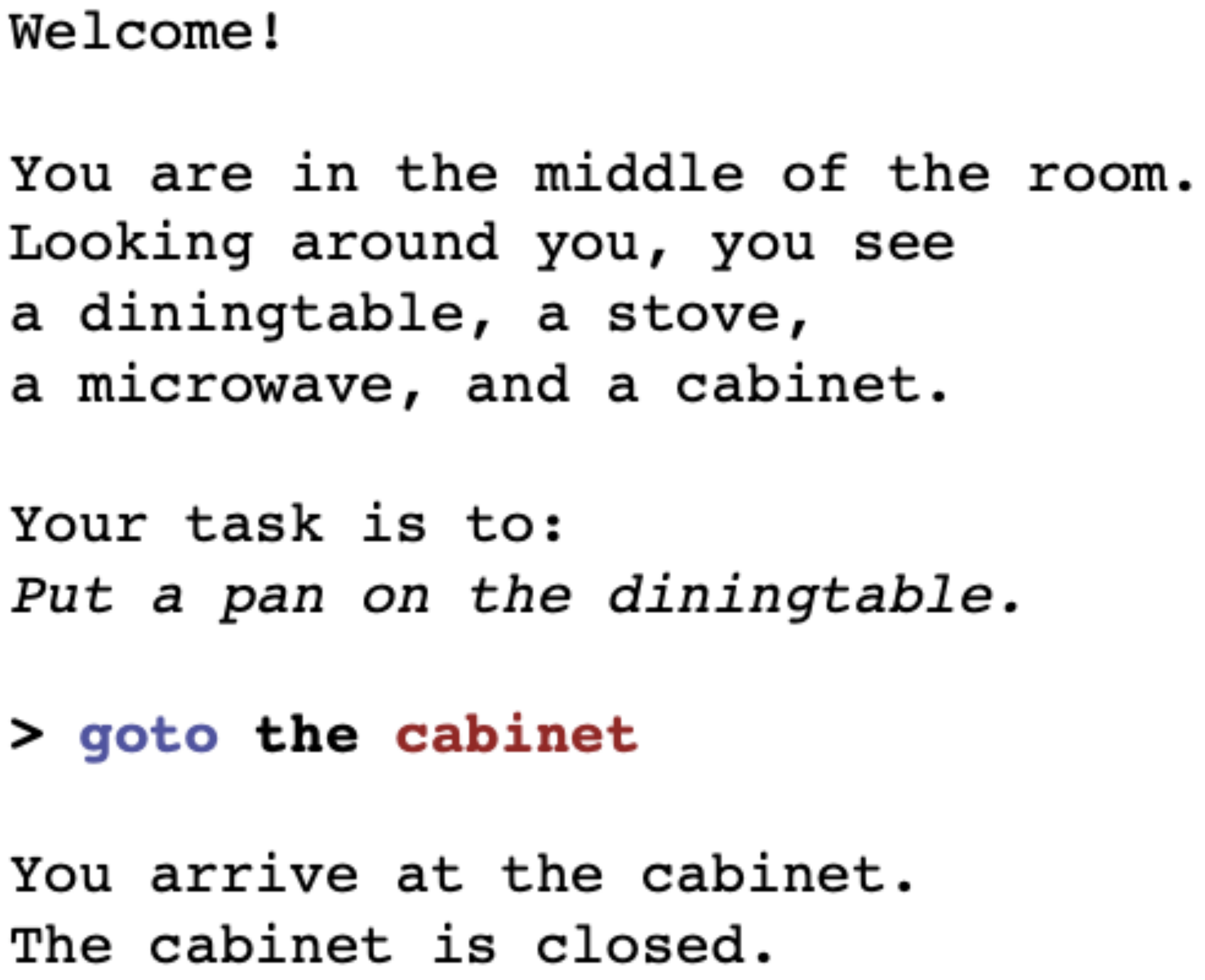
- Read goal specification and env feedback to perform long-horizon task in large procgen world
- Challenges
  - Partially observed, complex world
  - Very delayed, sparse reward

- Included modifications
  - Uniformly sample from subtasks Score, Gold, and Scout
  - Present goal text indicating which task to perform
  - Train single agent for all subtasks

Implementation

Included environments: ALFWorld

- Read goal, feedback, and scene description text to perform task in partially observed world
- Challenges
  - Partially observed world
  - Large language action space



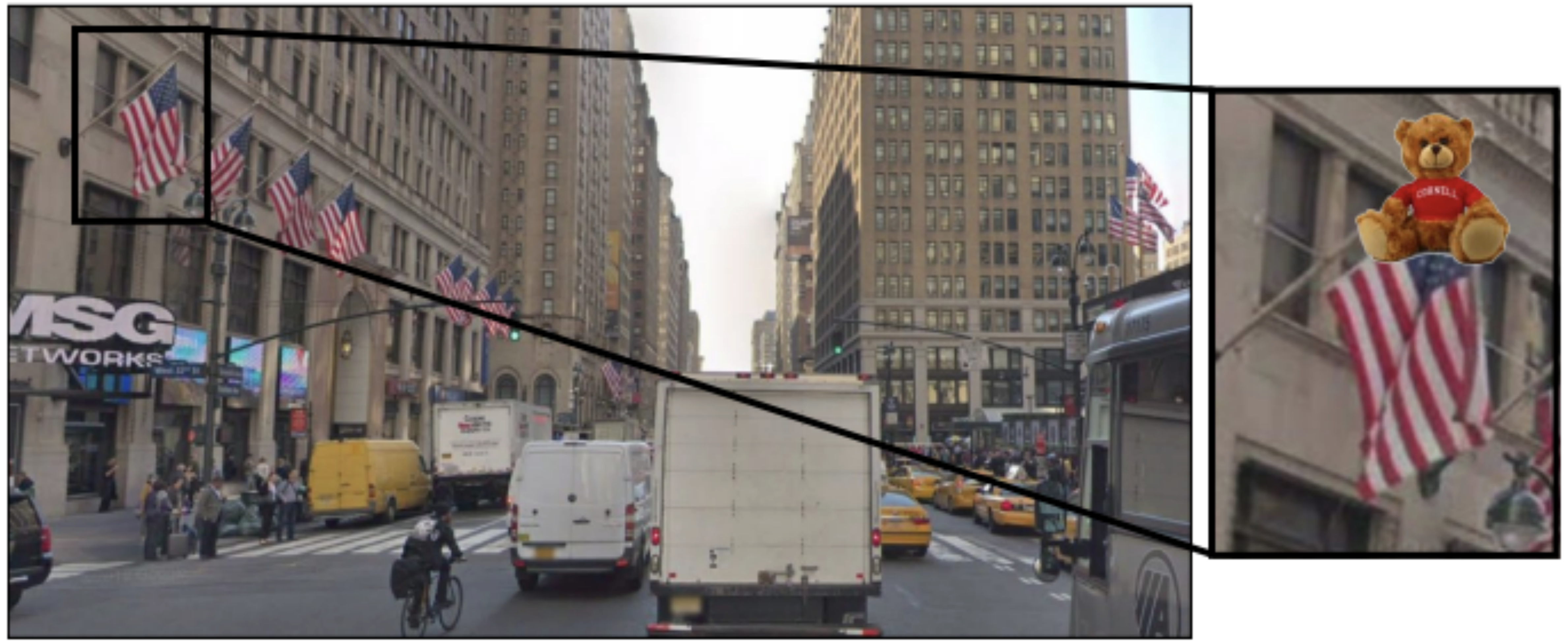
```
Welcome!

You are in the middle of the room.
Looking around you, you see
a diningtable, a stove,
a microwave, and a cabinet.

Your task is to:
Put a pan on the diningtable.

> goto the cabinet

You arrive at the cabinet.
The cabinet is closed.
```

ALFWorld: Aligning Text and Embodied Environments for Interactive Learning
Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, Matthew Hausknecht
ICLR 2021

Implementation

Included environments: SILGTouchdown

- Read natural language instruction to navigate panorama views from Google StreetView
- Challenges
  - Complex NL compositional instructions
  - Complex natural scenes



Turn and go with the flow of traffic. At the first traffic light turn left. Go past the next two traffic light, As you come to the third traffic light you will see a white building on your left with many American flags on it. Touchdown is sitting in the stars of the first flag.

Touchdown: Natural Language Navigation and Spatial Reasoning in Visual Street Environments
Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, Yoav Artzi
CVPR 2019

Implementation

Included environments: SILGTouchdown

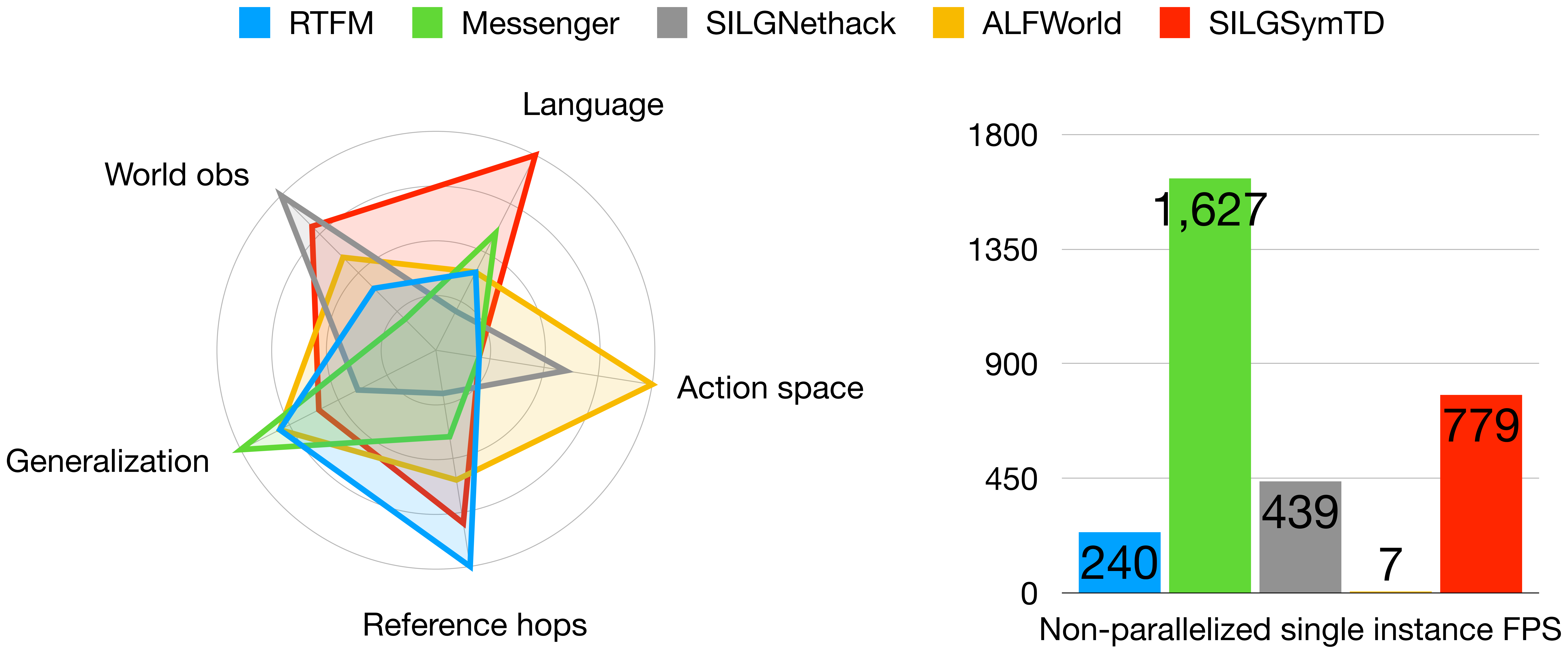
- Read natural language instruction to navigate panorama views from Google StreetView
- Challenges
  - Complex NL compositional instructions
  - Complex natural scenes

- Included modifications
  - Only consider navigation subtask by choosing direction of movement in panorama
  - SILGSymTouchdown
    - Apply segmentation model to obtain grid of object class IDs
  - SILGVisTouchdown
    - Top-k principal components of ResNet features of panorama
  - Manual variants
    - Agent must choose to stop at goal instead of auto-stopping

Touchdown: Natural Language Navigation and Spatial Reasoning in Visual Street Environments
Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, Yoav Artzi
CVPR 2019

Implementation

Environment comparison



Proposal

SILG: Situated Interactive Language Grounding Benchmark

- Combines unique generalization challenges
    - Complex scenes (SILGNethack, ALFWorld, SILGTouchdown)
    - New natural language references (Messenger, SILGTouchdown)
    - Large language action space (ALFWorld)
    - Multi-hop references/reasoning (RTFM, SILGTouchdown)
    - New entity dynamics (RTFM, Messenger)

Proposal

SILG: Situated Interactive Language Grounding Benchmark

- Combines unique generalization challenges
- Efficient environment for RL
    - Focus on symbolic envs with semantic symbols instead of raw visuals

Proposal

SILG: Situated Interactive Language Grounding Benchmark

- Combines unique generalization challenges
- Efficient environment for RL
- Easy to use
    - Shared OpenAI Gym interface
    - Included distributed RL framework via Torchbeast (just write policy model)

Proposal

SILG: Situated Interactive Language Grounding Benchmark

- Combines unique generalization challenges
- Efficient environment for RL
- Easy to use
- Goal
    - Quickly test new methods that generalize to diverse grounding challenges
    - Identify gaps in current tasks/setups for creation of new environments

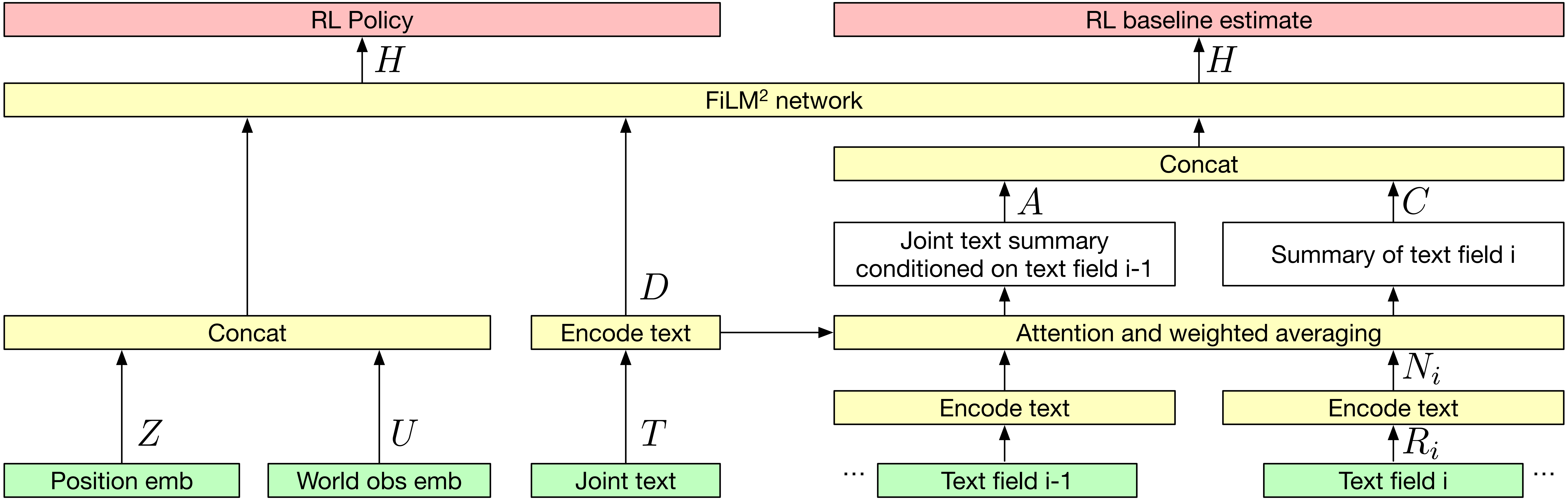
Experiments

What can we answer with SILG?

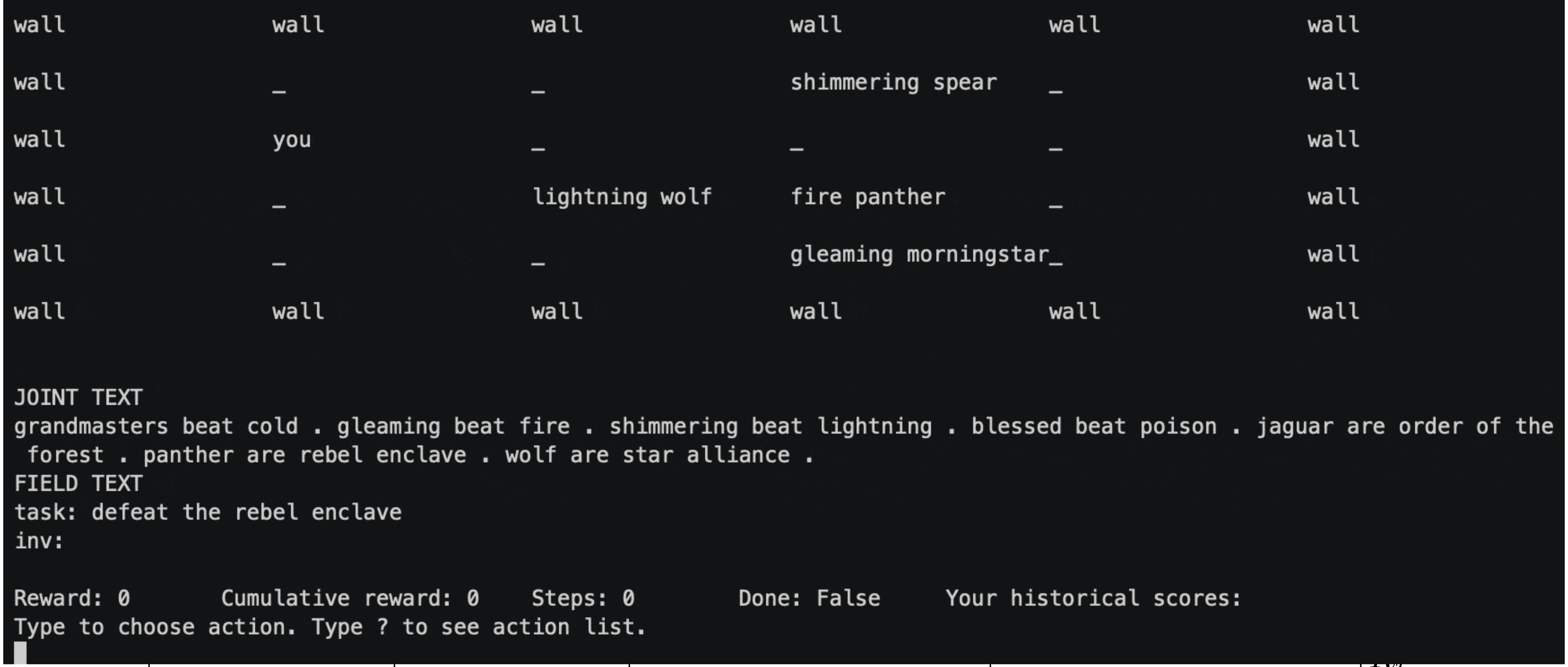
- Prior work used different methods for different envs
    - SOTA models vary in feature design, architecture, learning algorithms
- Can we use same feature+arch+learning alg combo across envs?
- What are some consistent findings across envs?

Experiments

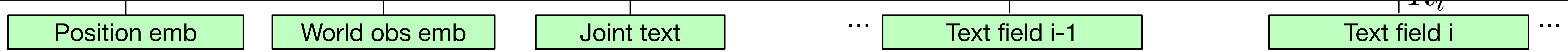
Situated Interactive Reader baseline (SIR)



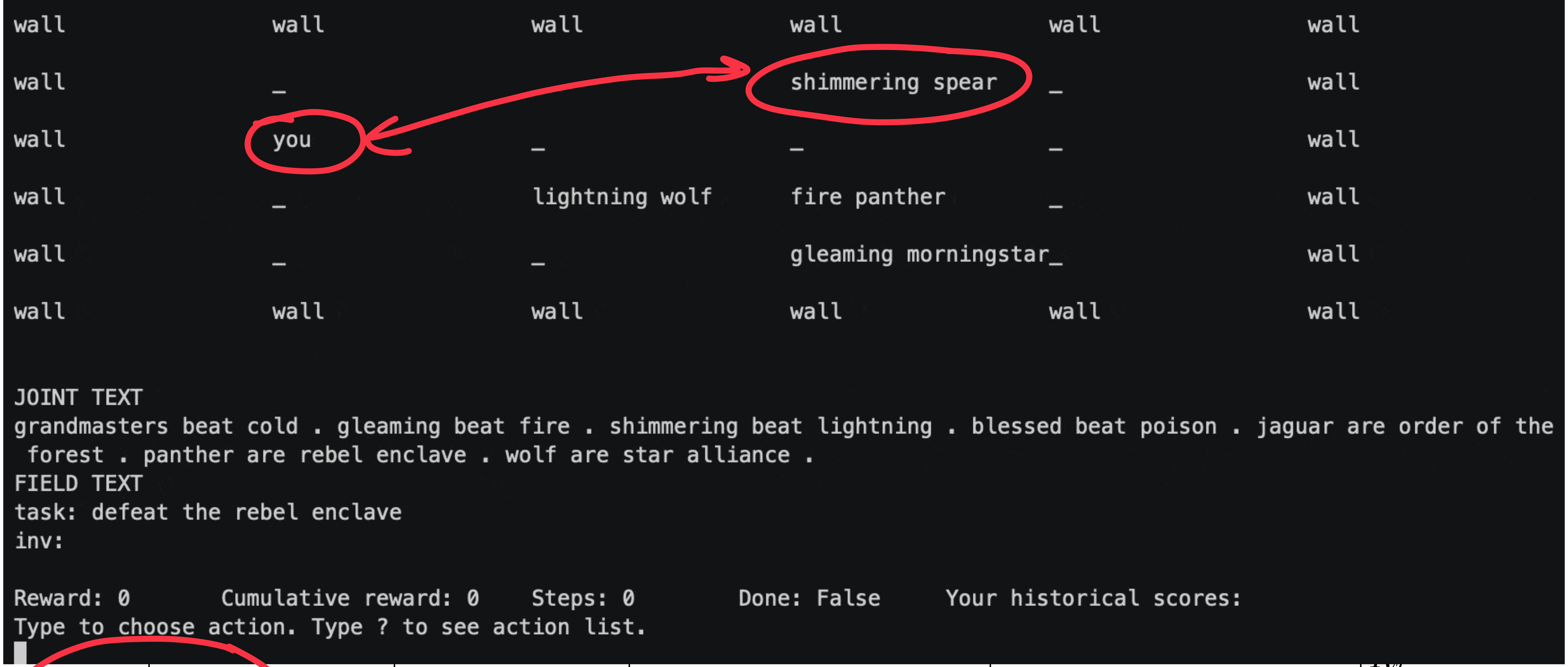
- Receives structured inputs



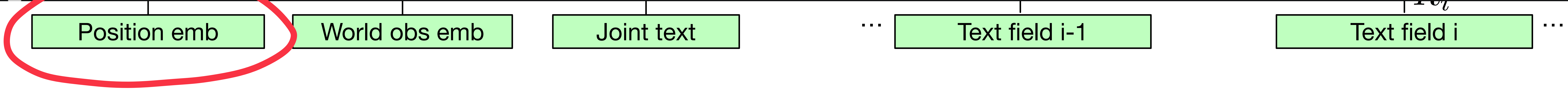
```
wall                wall                wall                wall                wall                wall

wall                _                   _                   shimmering spear    _                   wall

wall                you                 _                   _                   _                   wall

wall                _                   lightning wolf      fire panther        _                   wall

wall                _                   _                   gleaming morningstar_                   wall

wall                wall                wall                wall                wall                wall


JOINT TEXT
grandmasters beat cold . gleaming beat fire . shimmering beat lightning . blessed beat poison . jaguar are order of the
 forest . panther are rebel enclave . wolf are star alliance .
FIELD TEXT
task: defeat the rebel enclave
inv:

Reward: 0       Cumulative reward: 0    Steps: 0        Done: False     Your historical scores:
Type to choose action. Type ? to see action list.
```



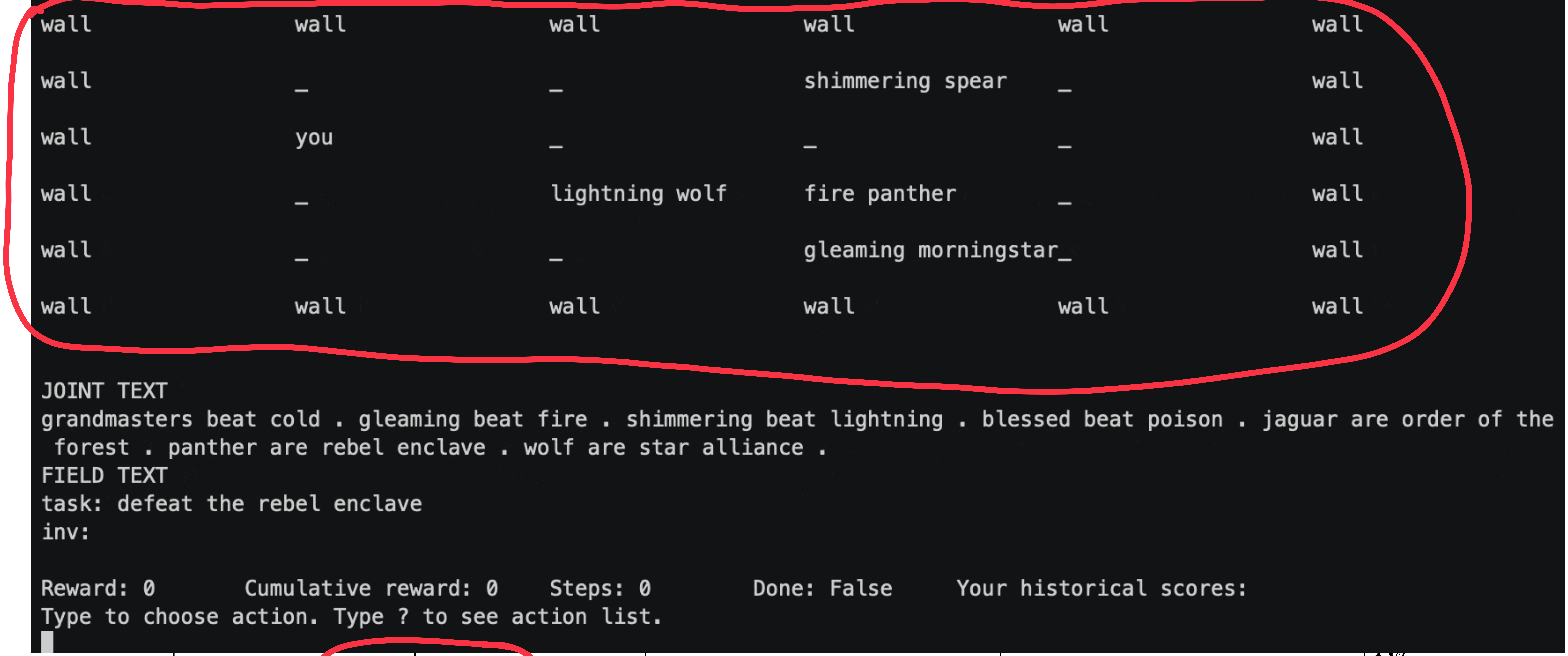
- RTFM as example



```
wall           wall           wall           wall               wall           wall

wall           _                             shimmering spear   _              wall

wall           you            _              _                  _              wall

wall           _              lightning wolf fire panther       _              wall

wall           _              _              gleaming morningstar_             wall

wall           wall           wall           wall               wall           wall


JOINT TEXT
grandmasters beat cold . gleaming beat fire . shimmering beat lightning . blessed beat poison . jaguar are order of the
 forest . panther are rebel enclave . wolf are star alliance .
FIELD TEXT
task: defeat the rebel enclave
inv:

Reward: 0       Cumulative reward: 0    Steps: 0        Done: False     Your historical scores:
Type to choose action. Type ? to see action list.
```



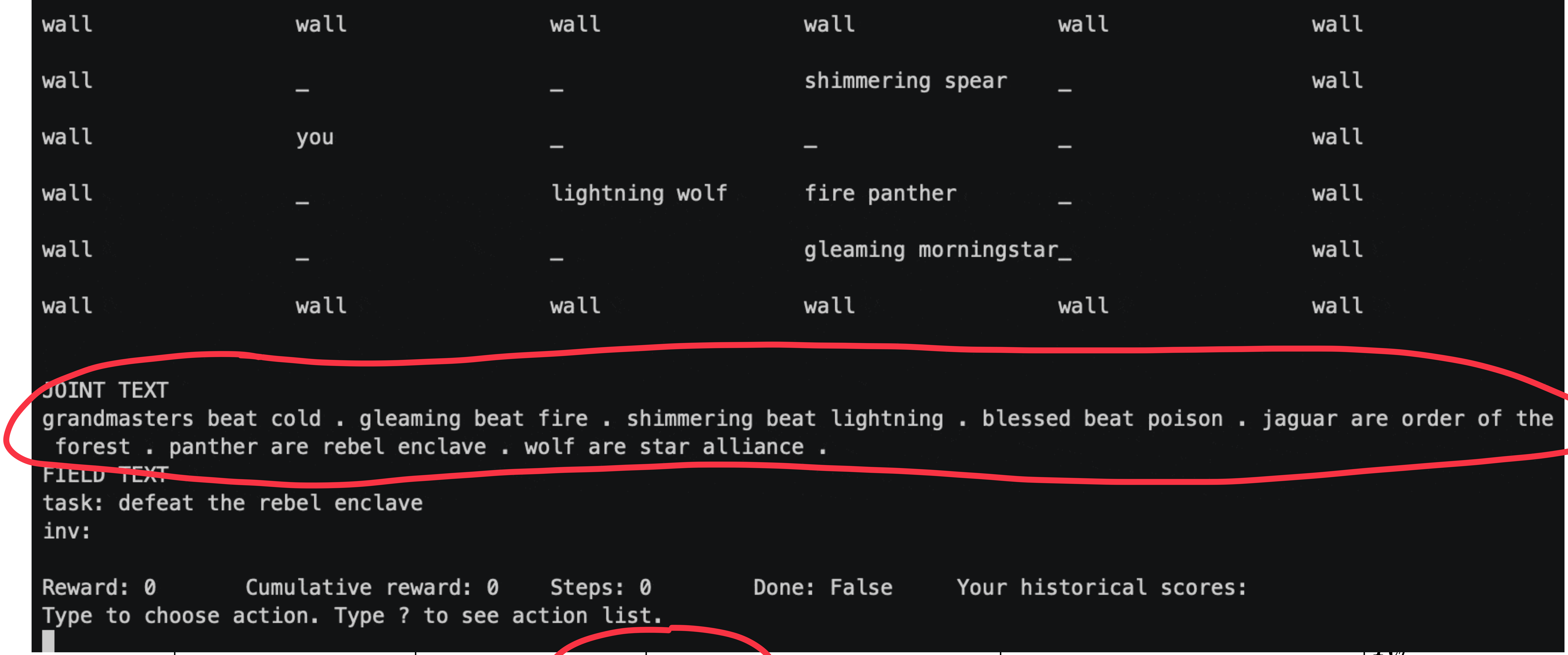
- XY distance of each cell relative to agent



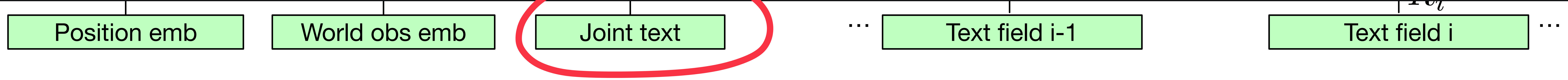
```
wall                wall                wall                wall                wall                wall

wall                _                   _                   shimmering spear    _                   wall

wall                you                 _                   _                   _                   wall

wall                _                   lightning wolf      fire panther        _                   wall

wall                _                   _                   gleaming morningstar_                   wall

wall                wall                wall                wall                wall                wall


JOINT TEXT
grandmasters beat cold . gleaming beat fire . shimmering beat lightning . blessed beat poison . jaguar are order of the
 forest . panther are rebel enclave . wolf are star alliance .
FIELD TEXT
task: defeat the rebel enclave
inv:

Reward: 0       Cumulative reward: 0    Steps: 0        Done: False     Your historical scores:
Type to choose action. Type ? to see action list.
```

Position emb
World obs emb
Joint text
…
Text field i-1
Text field i
…

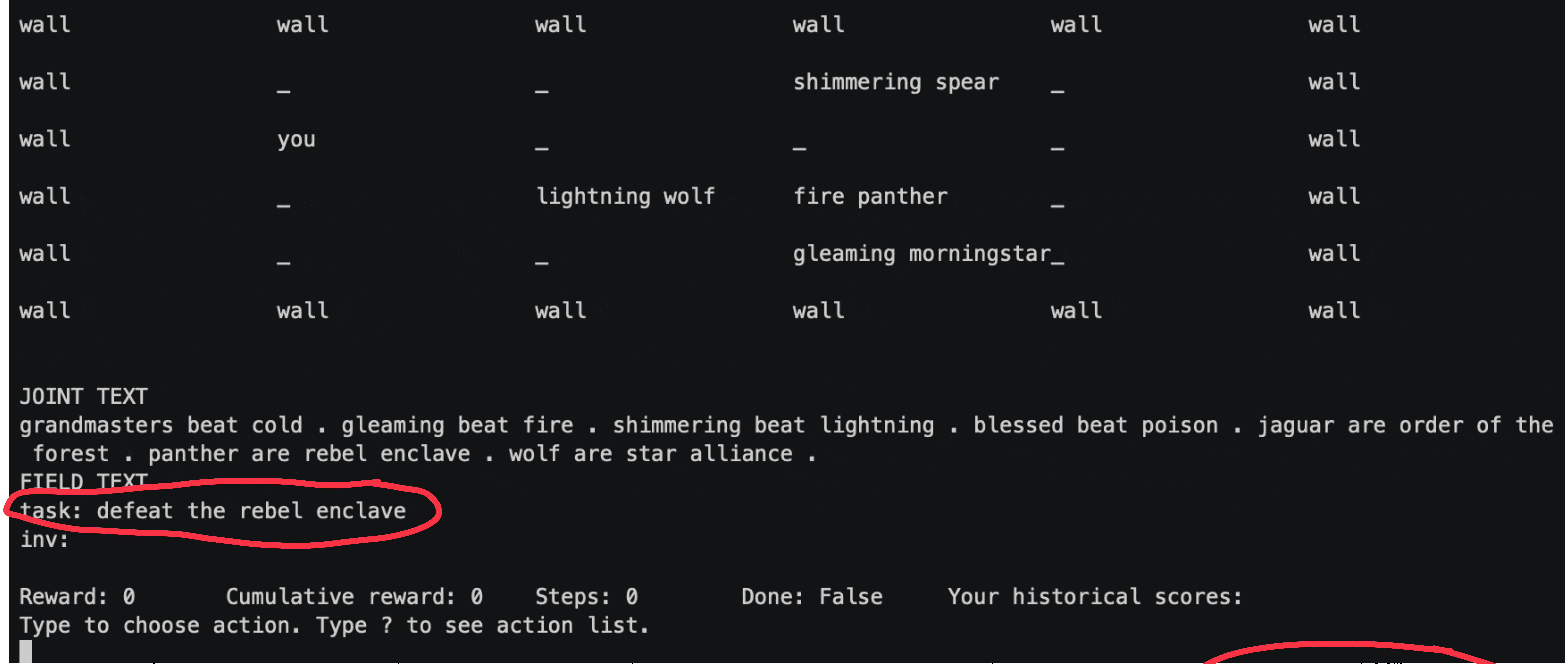
- Emb of entities in each cell



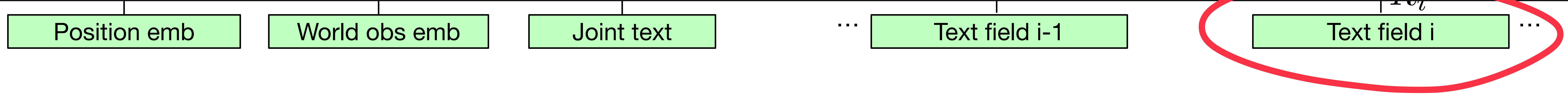
```
wall            wall            wall            wall            wall            wall

wall            _               _               shimmering spear _              wall

wall            you             _               _               _               wall

wall            _               lightning wolf  fire panther    _               wall

wall            _               _               gleaming morningstar_           wall

wall            wall            wall            wall            wall            wall


JOINT TEXT
grandmasters beat cold . gleaming beat fire . shimmering beat lightning . blessed beat poison . jaguar are order of the
 forest . panther are rebel enclave . wolf are star alliance .
FIELD TEXT
task: defeat the rebel enclave
inv:

Reward: 0       Cumulative reward: 0    Steps: 0        Done: False     Your historical scores:
Type to choose action. Type ? to see action list.
```



- Joint text to attend over (RTFM wiki in this case)



```
wall                wall                wall                wall                wall                wall

wall                _                   _                   shimmering spear    _                   wall

wall                you                 _                   _                   _                   wall

wall                _                   lightning wolf      fire panther        _                   wall

wall                _                   _                   gleaming morningstar_                   wall

wall                wall                wall                wall                wall                wall


JOINT TEXT
grandmasters beat cold . gleaming beat fire . shimmering beat lightning . blessed beat poison . jaguar are order of the
 forest . panther are rebel enclave . wolf are star alliance .
FIELD TEXT
task: defeat the rebel enclave
inv:

Reward: 0       Cumulative reward: 0    Steps: 0        Done: False     Your historical scores:
Type to choose action. Type ? to see action list.
```



- Env-specific text fields (goal text, inventory text in this case)

Structured input for each env

Environment	World observation emb	Joint text	Text fields	Action layer
RTFM	Sum of embedding of entity name in grid	Wiki of high level rules	Goal, inventory	MLP over movement directions
Messenger	Embedding of entity IDs in grid	Concat text fields	Clauses in manual	MLP over movement directions
SILGNethack	Embeddings of entity IDs in grid	Concat text fields	Goal, feedback	MLP over NetHack actions
ALFWorld	Sum of embeddings of entity names in scene	Concat text fields	Goal, feedback, history of past actions and feedback	Rank over RNN encoding of valid language actions
SILGTouchdown	Embeddings of entity IDs in panorama grid (or PCA of ResNet features)	Instruction	Instruction	Index into representation slice corresponding to navigation direction

Structured input for each env

Grid worlds

Environment	World observation emb	Joint text	Text fields	Action layer
RTFM	Sum of embedding of entity name in grid	Wiki of high level rules	Goal, inventory	MLP over movement directions
Messenger	Embedding of entity IDs in grid	Concat text fields	Clauses in manual	MLP over movement directions
SILGNethack	Embeddings of entity IDs in grid	Concat text fields	Goal, feedback	MLP over NetHack actions
ALFWorld	Sum of embeddings of entity names in scene	Concat text fields	Goal, feedback, history of past actions and feedback	Rank over RNN encoding of valid language actions
SILGTouchdown	Embeddings of entity IDs in panorama grid (or PCA of ResNet features)	Instruction	Instruction	Index into representation slice corresponding to navigation direction

Structured input for each env

Environment	World observation emb	Joint text	Text fields	Action layer
RTFM	Sum of embedding of entity name in grid	Wiki of high level rules	Goal, inventory	MLP over movement directions
Messenger	Embedding of entity IDs in grid	Concat text fields	Clauses in manual	MLP over movement directions
SILGNethack	Embeddings of entity IDs in grid	Concat text fields	Goal, feedback	MLP over NetHack actions
ALFWorld	Sum of embeddings of entity names in scene	Concat text fields	Goal, feedback, history of past actions and feedback	Rank over RNN encoding of valid language actions
SILGTouchdown	Embeddings of entity IDs in panorama grid (or PCA of ResNet features)	Instruction	Instruction	Index into representation slice corresponding to navigation direction

Structured input for each env

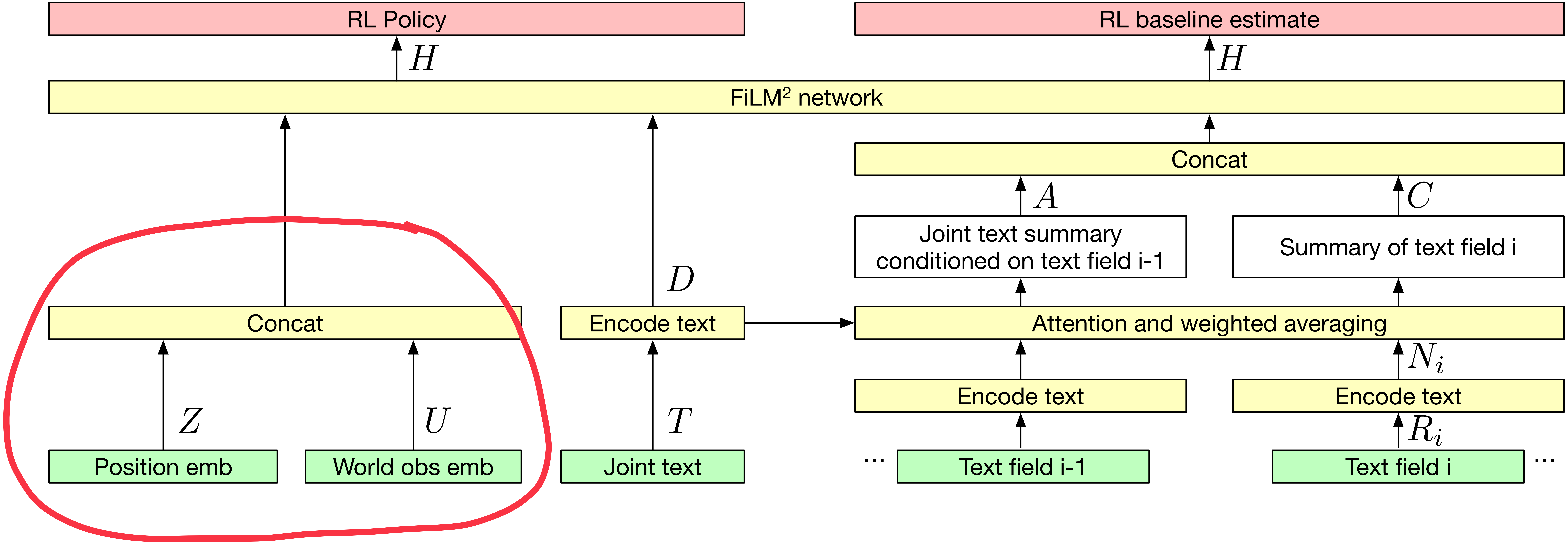
Long-horizon navigation

W/ real scenes + real language

Environment	World observation emb	Joint text	Text fields	Action layer
RTFM	Sum of embedding of entity name in grid	Wiki of high level rules	Goal, inventory	MLP over movement directions
Messenger	Embedding of entity IDs in grid	Concat text fields	Clauses in manual	MLP over movement directions
SILGNethack	Embeddings of entity IDs in grid	Concat text fields	Goal, feedback	MLP over NetHack actions
ALFWorld	Sum of embeddings of entity names in scene	Concat text fields	Goal, feedback, history of past actions and feedback	Rank over RNN encoding of valid language actions
SILGTouchdown	Embeddings of entity IDs in panorama grid (or PCA of ResNet features)	Instruction	Instruction	Index into representation slice corresponding to navigation direction

Experiments

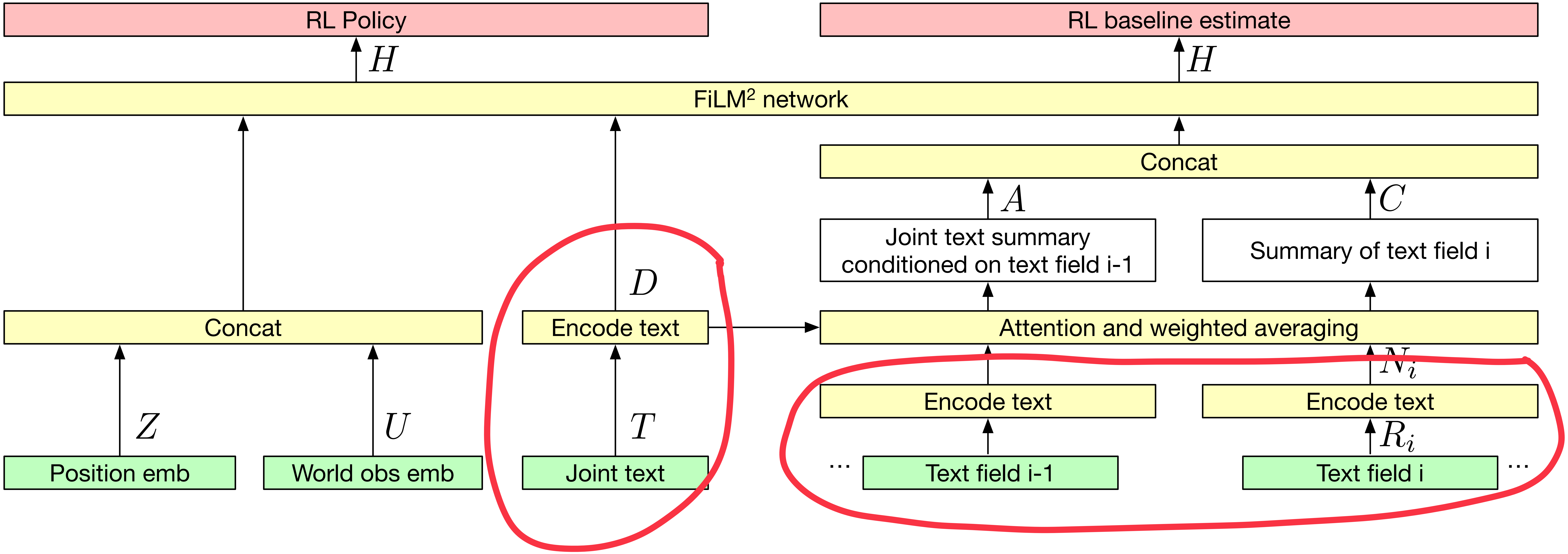
Situated Interactive Reader baseline (SIR)



- Concat position emb and world obs emb

Experiments

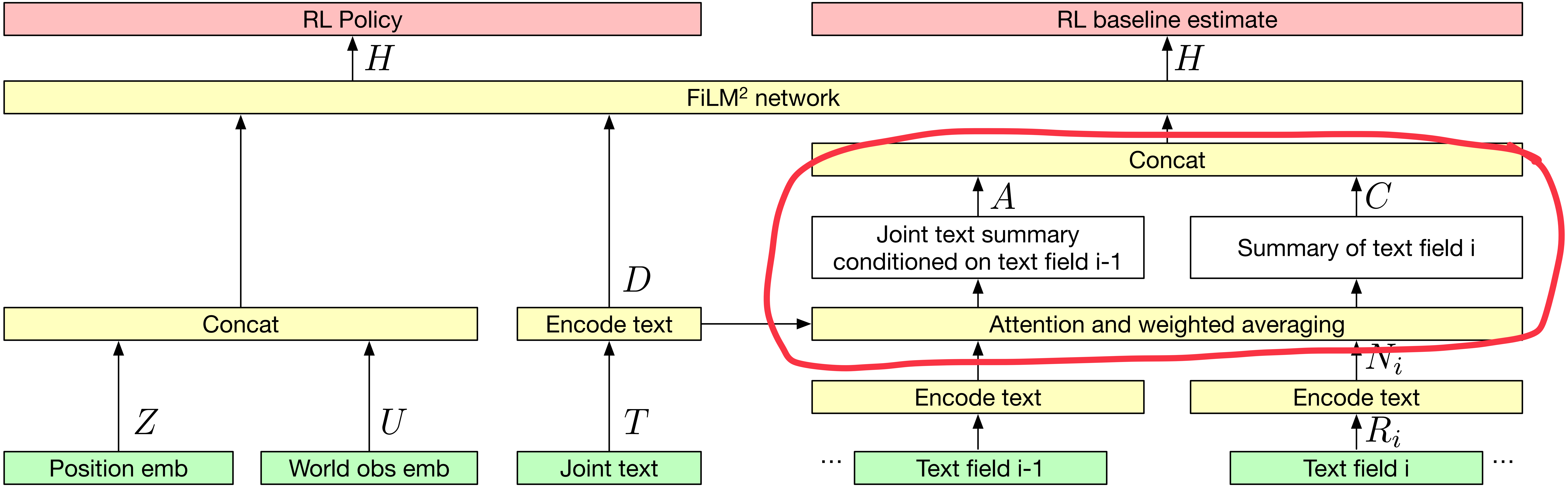
Situated Interactive Reader baseline (SIR)



- Encode text fields with LSTMs

Experiments

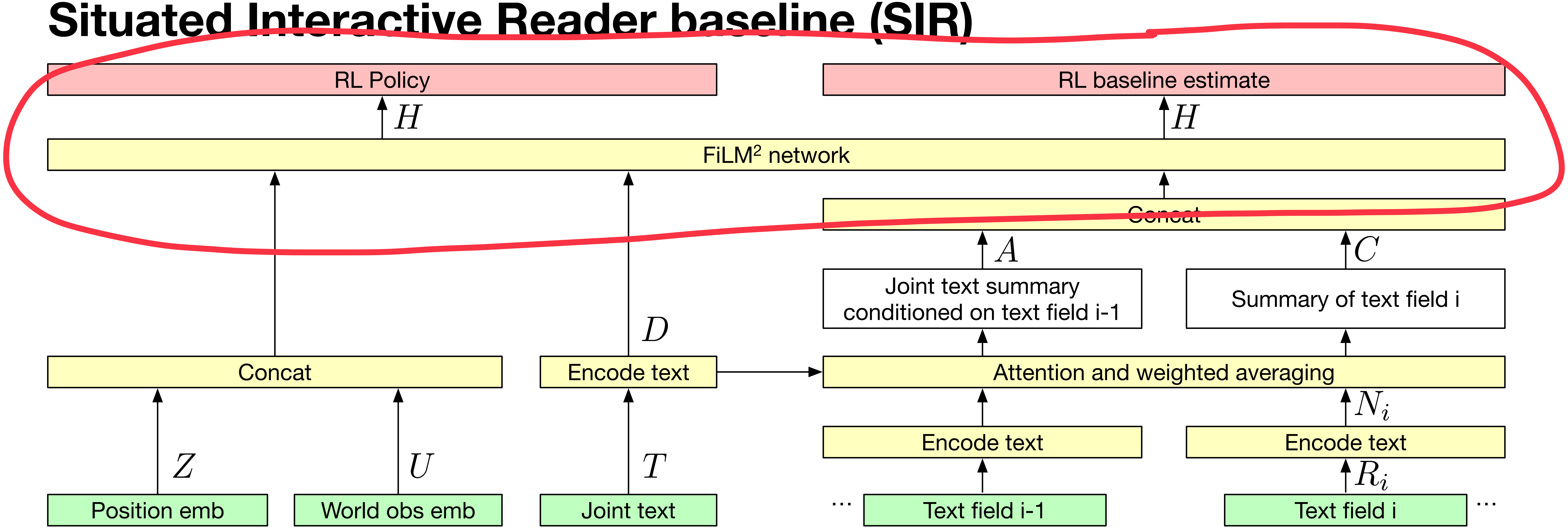
Situated Interactive Reader baseline (SIR)



- For each text field, compute attention over joint text and self-attention
- Combine via weighted sum

Experiments

Situated Interactive Reader baseline (SIR)



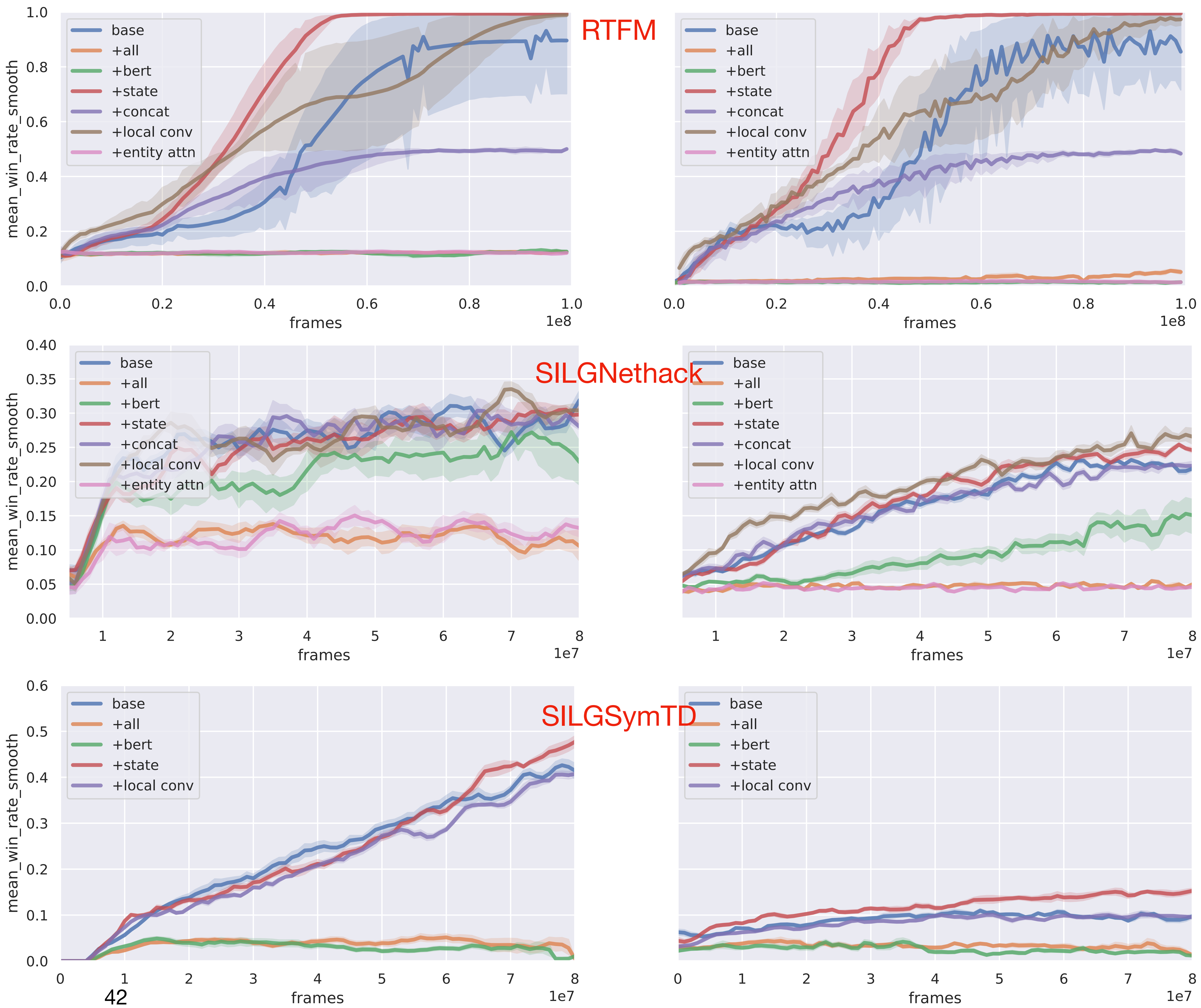
- Learn codependent representation with FiLM2
- Estimate baseline and distribution over actions

Results of SIR + enhancements

Env	Generalization evaluation	Best enhancement
RTFM	New scenes New rules	+state tracking
Messenger	New rules New NL references	+state tracking +local conv +entity attn +BERT
SILGNethack	New scenes	+local conv
ALFWorld	New instructions	+state tracking
ALFWorld +new scenes	New instructions New scenes	+state tracking
SILGSymTD	New instructions New scenes	+state tracking

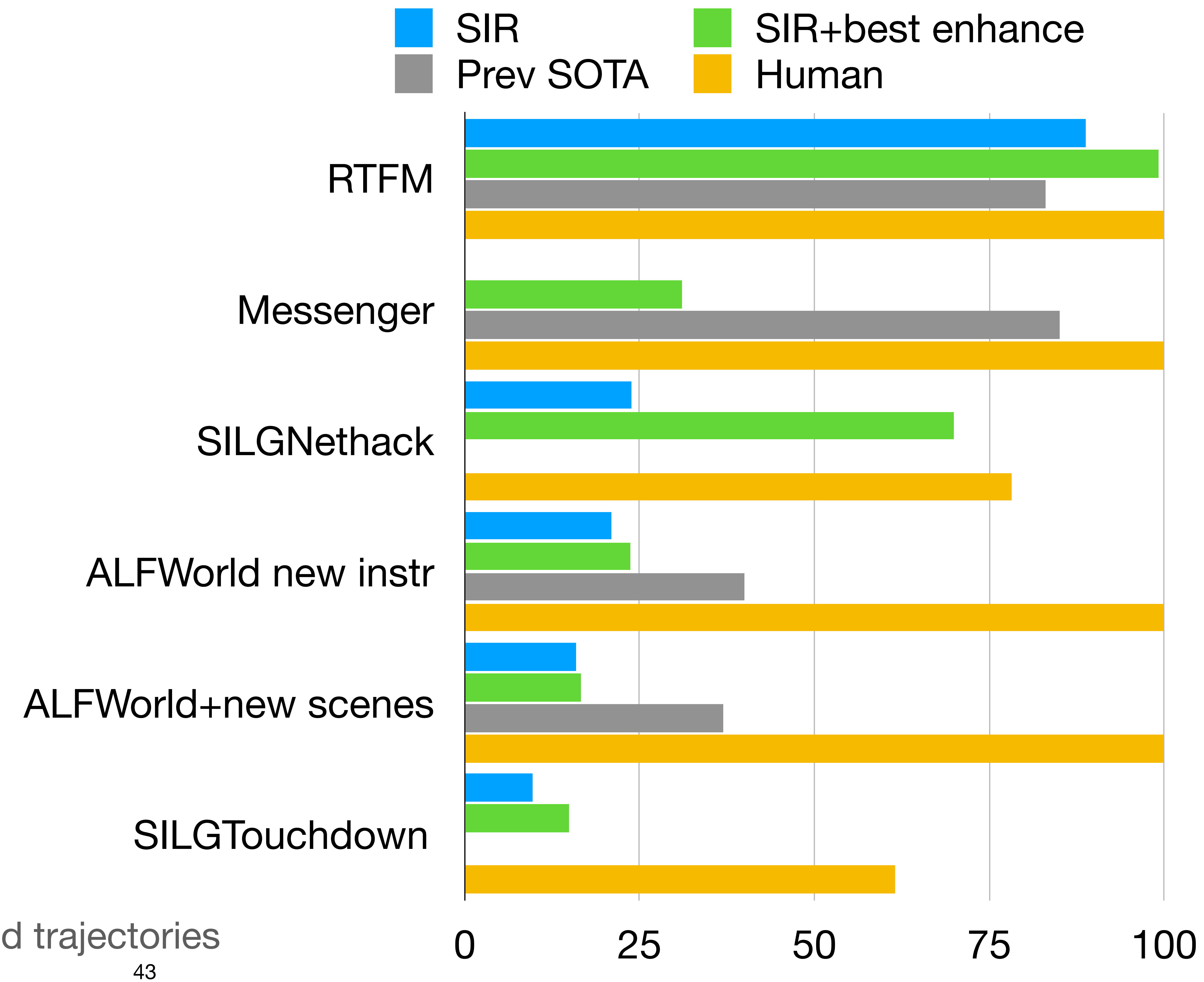
Results of SIR + enhancements

Env	Generalization evaluation	Best enhancement
RTFM	New scenes New rules	+state tracking
Messenger	New rules New NL references	+state tracking +local conv +entity attn +BERT
SILGNethack	New scenes	+local conv
ALFWorld	New instructions	+state tracking
ALFWorld +new scenes	New instructions New scenes	+state tracking
SILGSymTD	New instructions New scenes	+state tracking



Results of SIR + enhancements

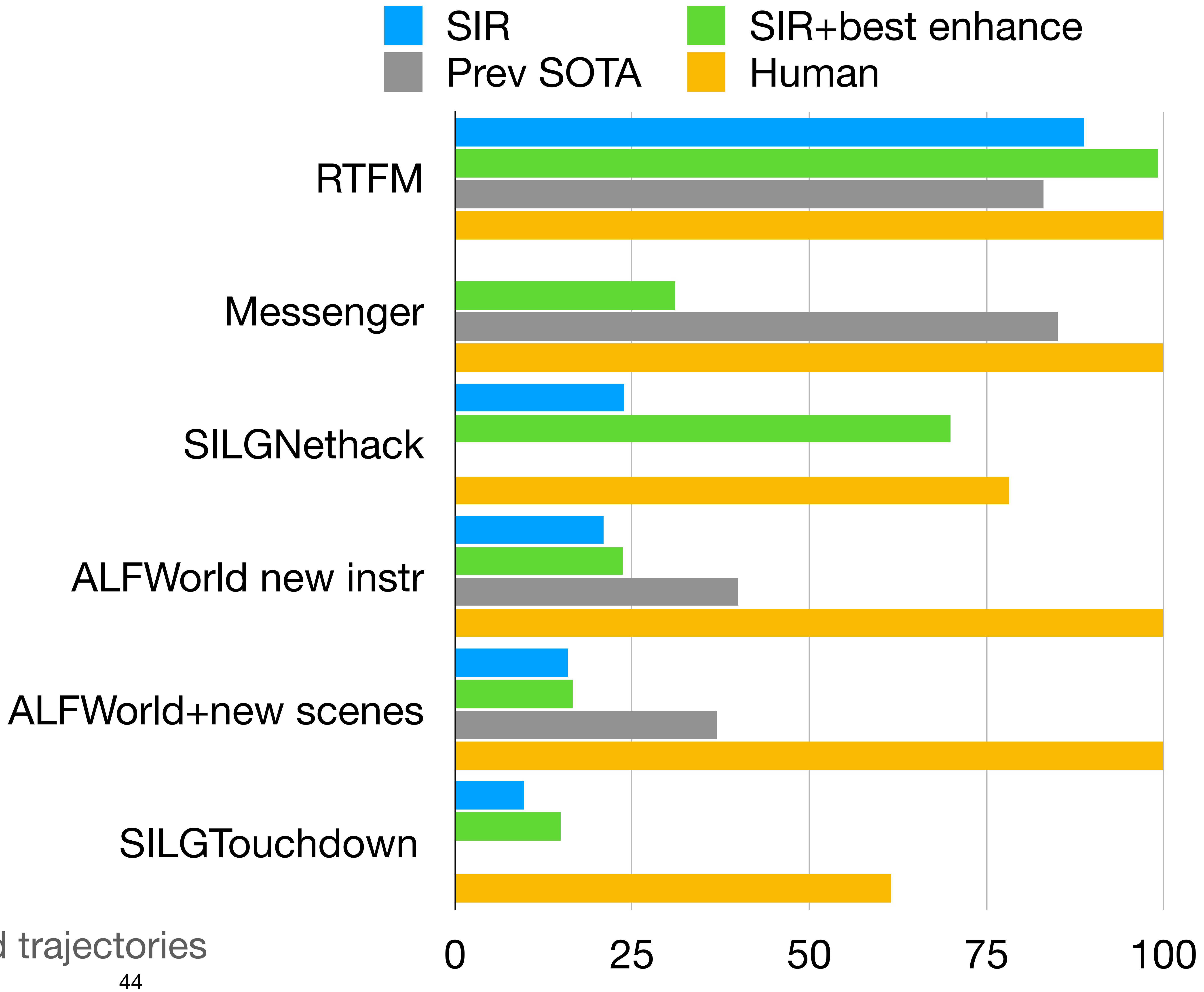
Env	Generalization evaluation	Best enhancement
RTFM	New scenes New rules	+state tracking
Messenger	New rules New NL references	+state tracking +local conv +entity attn +BERT
SILGNethack	New scenes	+local conv
ALFWorld	New instructions	+state tracking
ALFWorld +new scenes	New instructions New scenes	+state tracking
SILGSymTD	New instructions New scenes	+state tracking



*ALFWorld Prev SOTA is Imitation learning w/ labeled trajectories

Results of SIR + enhancements

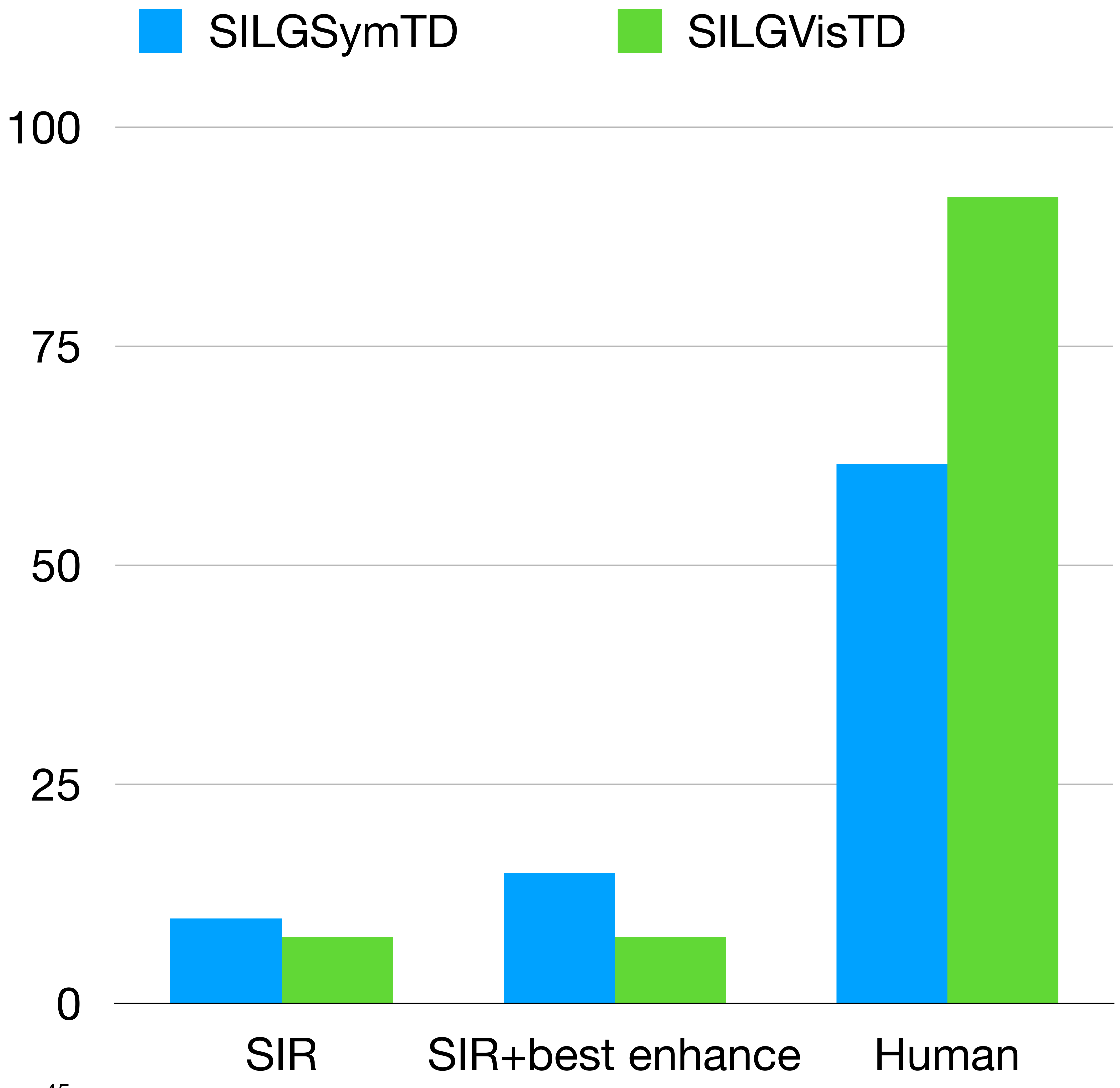
- Can train one arch across envs
- Consistent findings
  - Lexical generalization to new NL surprisingly difficult when grounding to new rules
    - Need better ways to incorporate pretrained LMs
  - State-tracking over pre-policy representation consistently helpful
  - Separating text fields for structured attention helps across envs than concatenating text for self-attention



*ALFWorld Prev SOTA is Imitation learning w/ labeled trajectories

Results of SIR + enhancements

- Is training on symbolic variant useful for achieving visual end-task?
  - Shridhar et al: training on text ALFWorld transfers to 3D ALFRED
  - Similarly here, training on SymTD transferrable to VisTD
    - Apply same segmentation model to VisTD, compute policy w/ SymTD model
    - Easier to RL in Sym (segmentation map of class IDs) vs. Vis (PCA of ResNet features) despite lower human upper bound



Summary

SILG: Situated Interactive Language Grounding Benchmark

- Combines unique generalization challenges
- Efficient environment for RL
- Easy to use
- Goal
    - Quickly test new methods that generalize to diverse grounding challenges
    - Identify gaps in current tasks/setups for creation of new environments

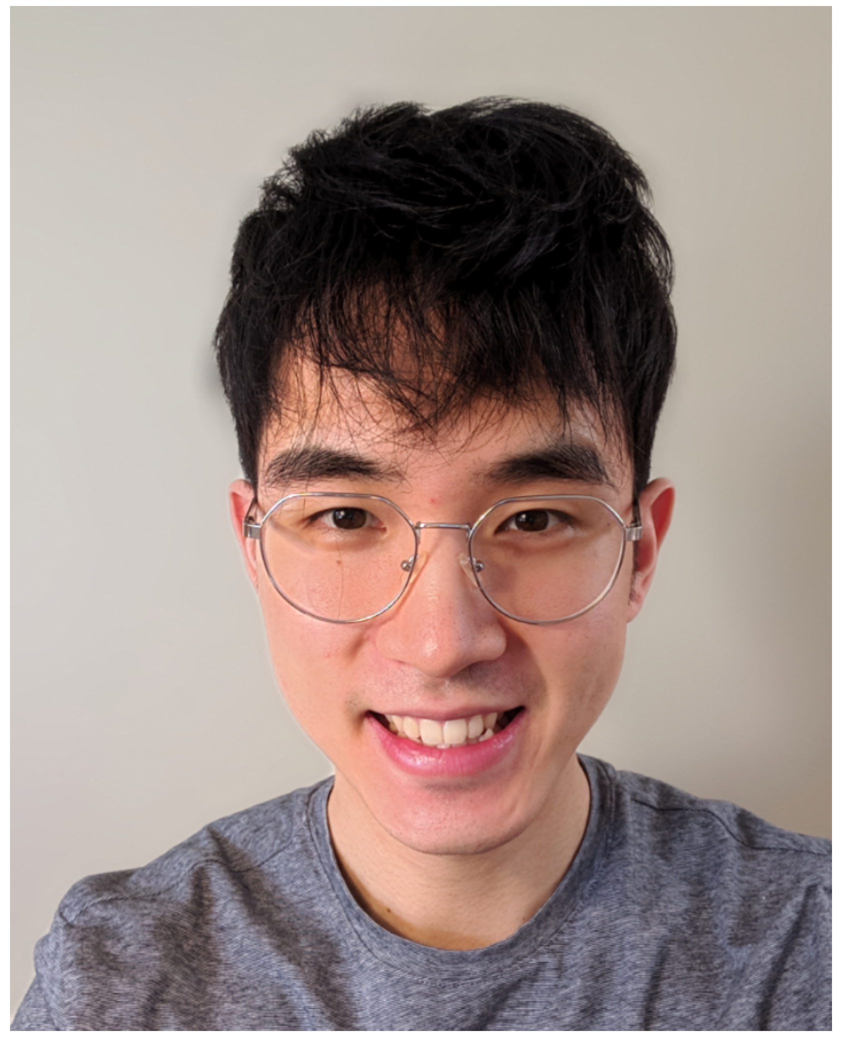
Release

https://github.com/vzhong/silg

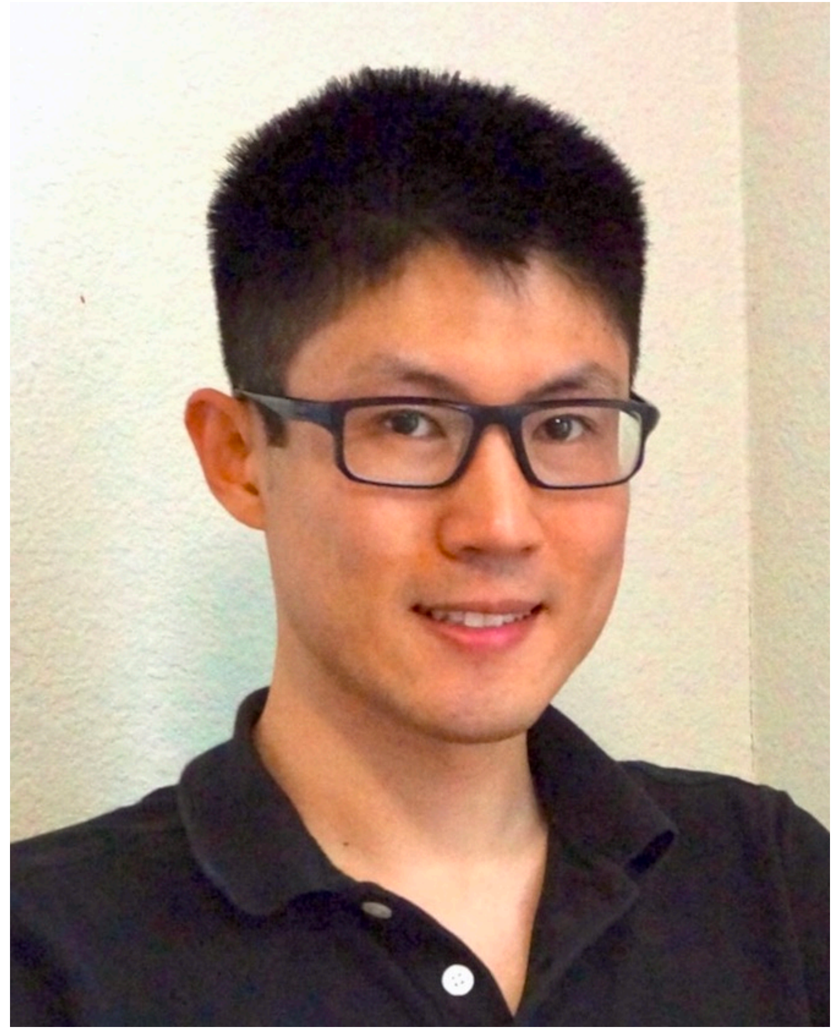
- Source code for SILG + SIR
- Docker image for installation
- Script to play envs yourself
- Instructions to add new envs!

Thank you!

Victor Zhong
University of Washington
Facebook AI Research



Austin W. Hanjie
Princeton University



Sida I. Wang
Facebook AI Research

Karthik Narasimhan
Princeton University

Luke Zettlemoyer
University of Washington
Facebook AI Research

UWNLP

PRINCETON NLP

FACEBOOK AI