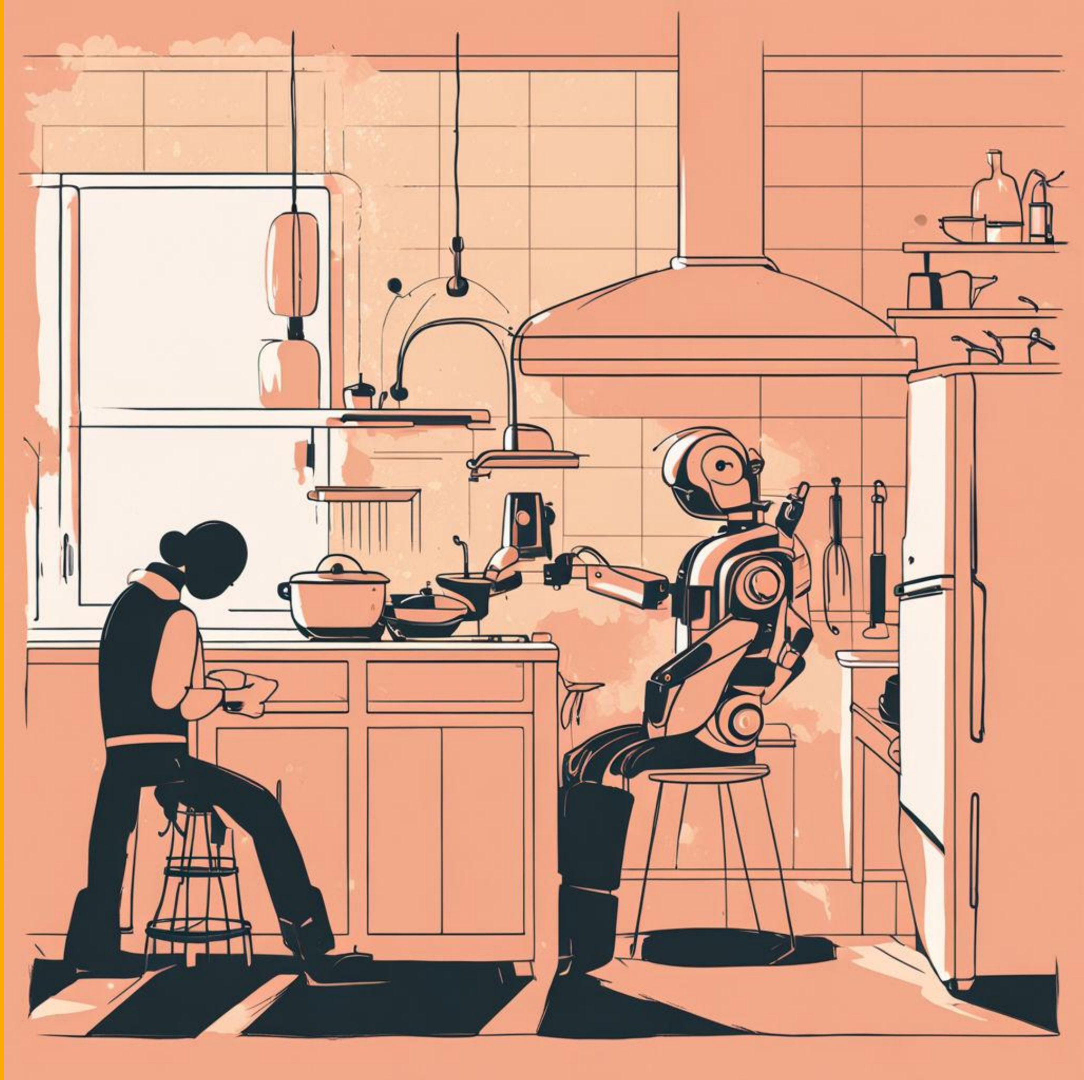
Learning to Model the World with Language

Jessy Lin, Yuqing Du, Olivia Watkins, Danijar Hafner,
Pieter Abbeel, Dan Klein, Anca Dragan

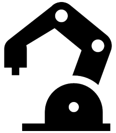
Berkeley
UNIVERSITY OF CALIFORNIA

Motivation: Interactive Embodied Agents



How do we develop agents that can communicate naturally with humans in the real world?

Hand me the cup.



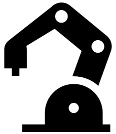
Here's how the coffee machine works: …

We're out of milk.

I already vacuumed the living room.

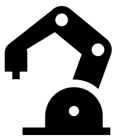
Instruction Following

"Move the cup to the table."



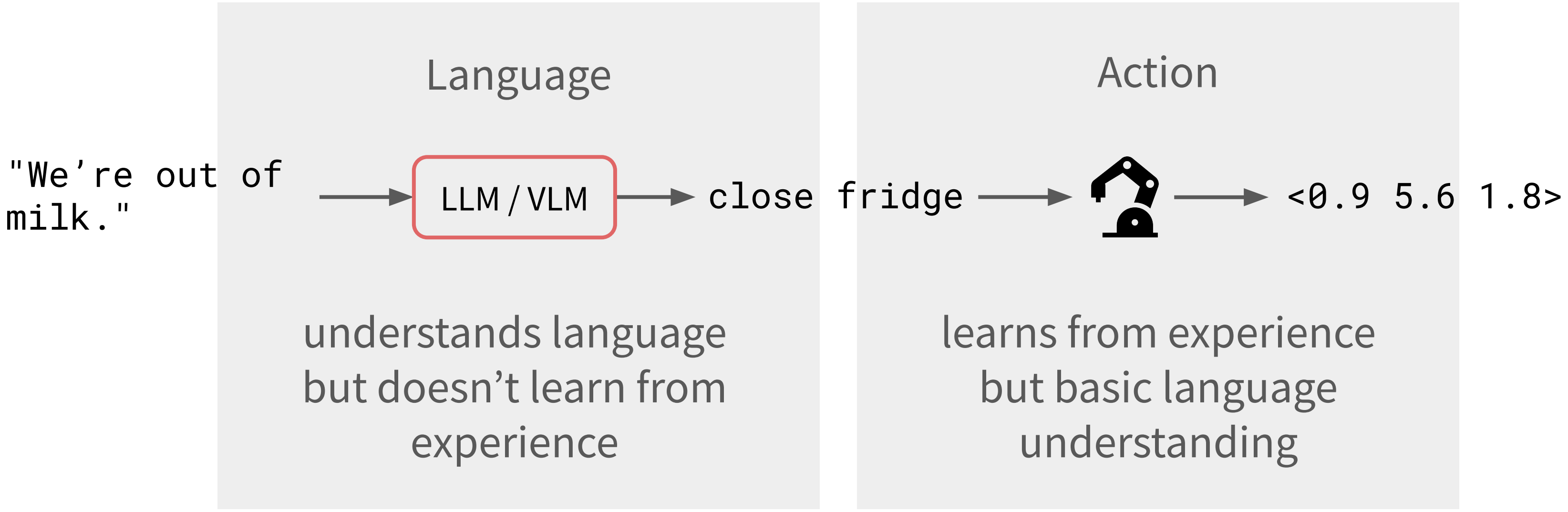
Action <0.9 5.6 1.8>

Instruction Following

"We're out of milk." →  → ?

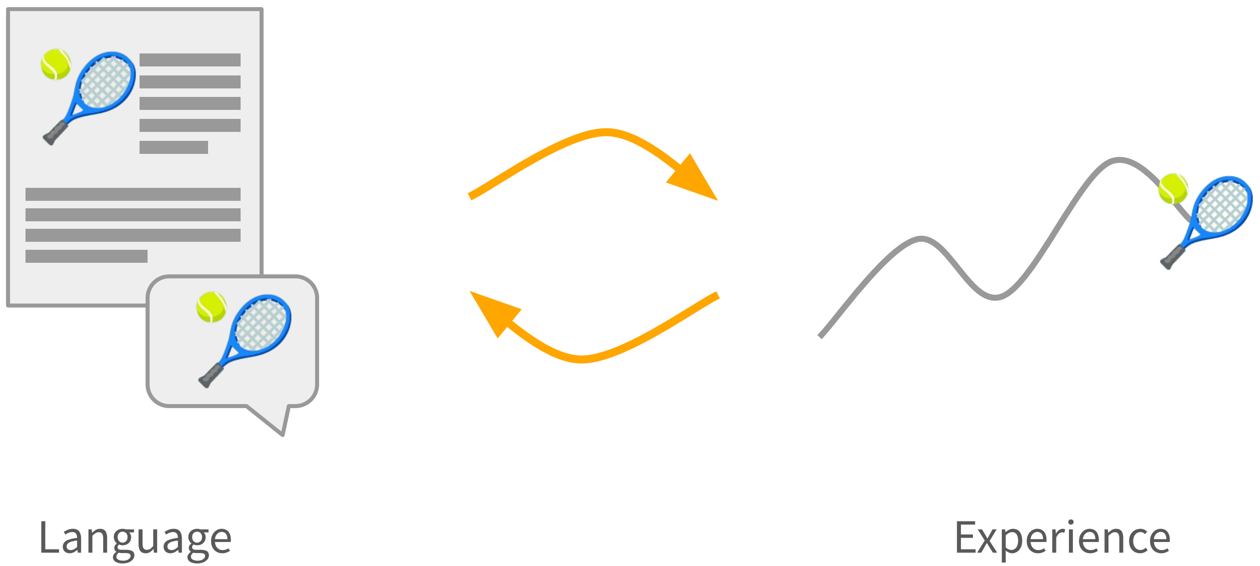
Interactive Embodied Agents

...with large language models or vision-language models?

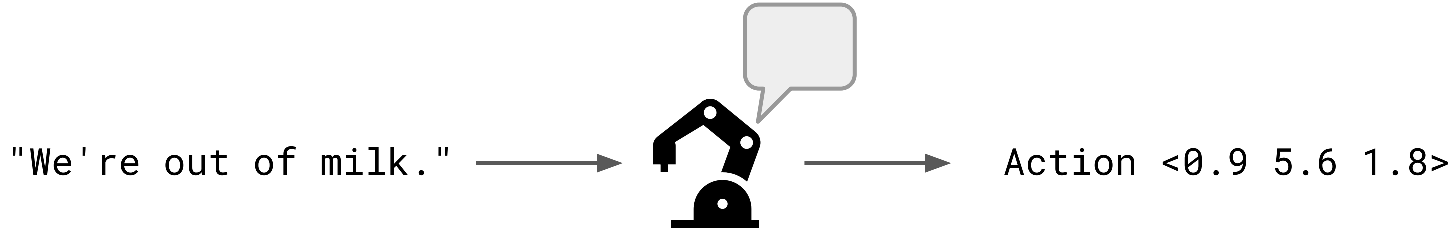


Interactive Embodied Agents

…with large language models or vision-language models?



Interactive Embodied Agents



How do we develop interactive embodied agents that can communicate naturally with humans?

Hand me the cup.

Here's how the coffee machine works: …

We're out of milk.

I already vacuumed the living room.

Key Idea

Language in the world can be understood as information that **helps agents predict the future** –

what will be observed, how the world will behave, and which situations will be rewarded.

Language beyond instructions as *world modeling*

The refrigerator is behind you.

The bread is being microwaved, it'll be done in 20 seconds.

If you press the black button, the microwave will open.

State

Future observations

Dynamics

→ right → right

Language beyond instructions as *world modeling*

The refrigerator is behind you.

The bread is being microwaved, it'll be done in 20 seconds.

If you press the black button, the microwave will open.

State

Future observations

Dynamics

t=0 … t=20

Language beyond instructions as *world modeling*

The refrigerator is behind you.

State

The bread is being microwaved, it'll be done in 20 seconds.

Future observations

If you press the black button, the microwave will open.

Dynamics

PUSH

Instruction following as *world modeling*

`Microwave the bread.`

Instructions

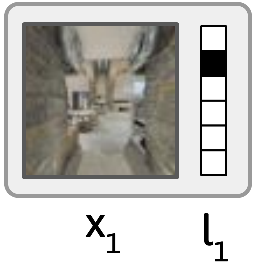
...

Key Idea

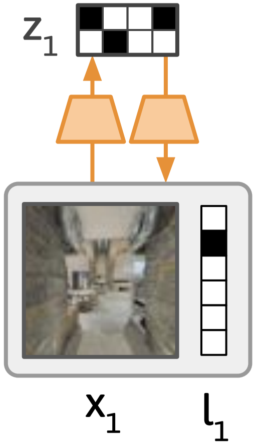
Language in the world can be understood as information that **helps agents predict the future** — suggesting a unifying self-supervised prediction objective.

Dynalang

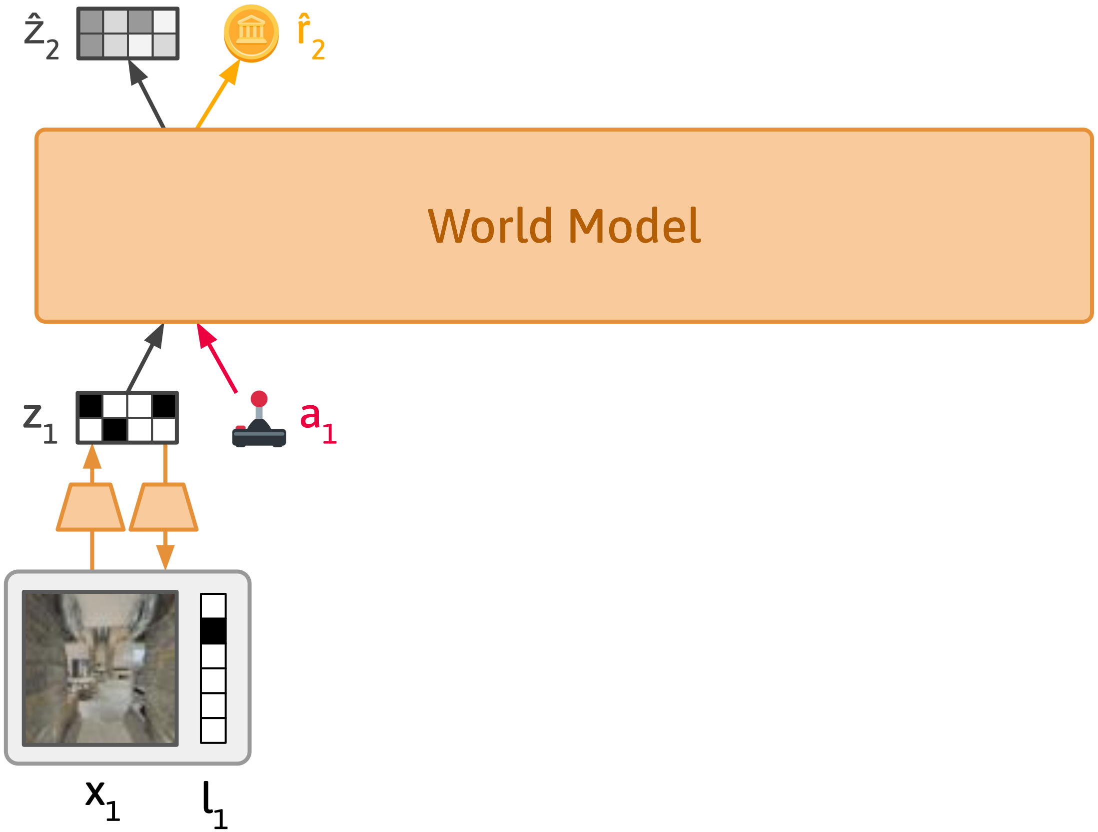
Dynalang



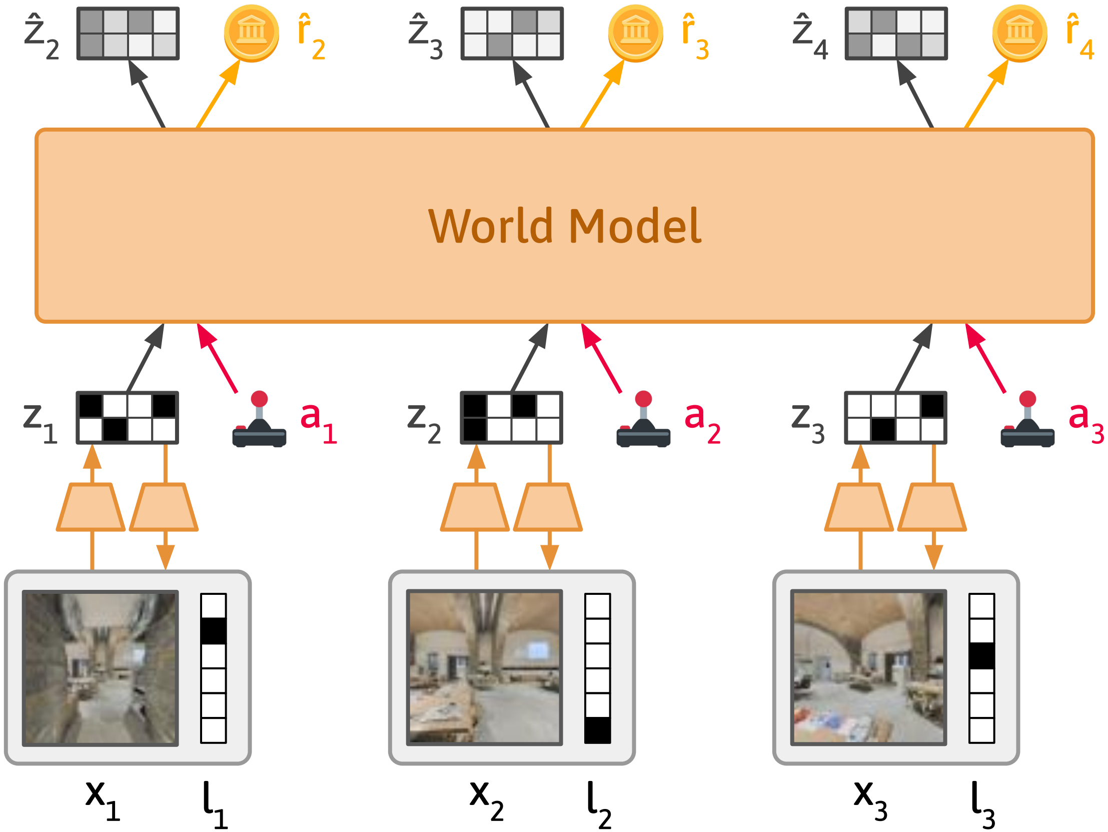
Dynalang



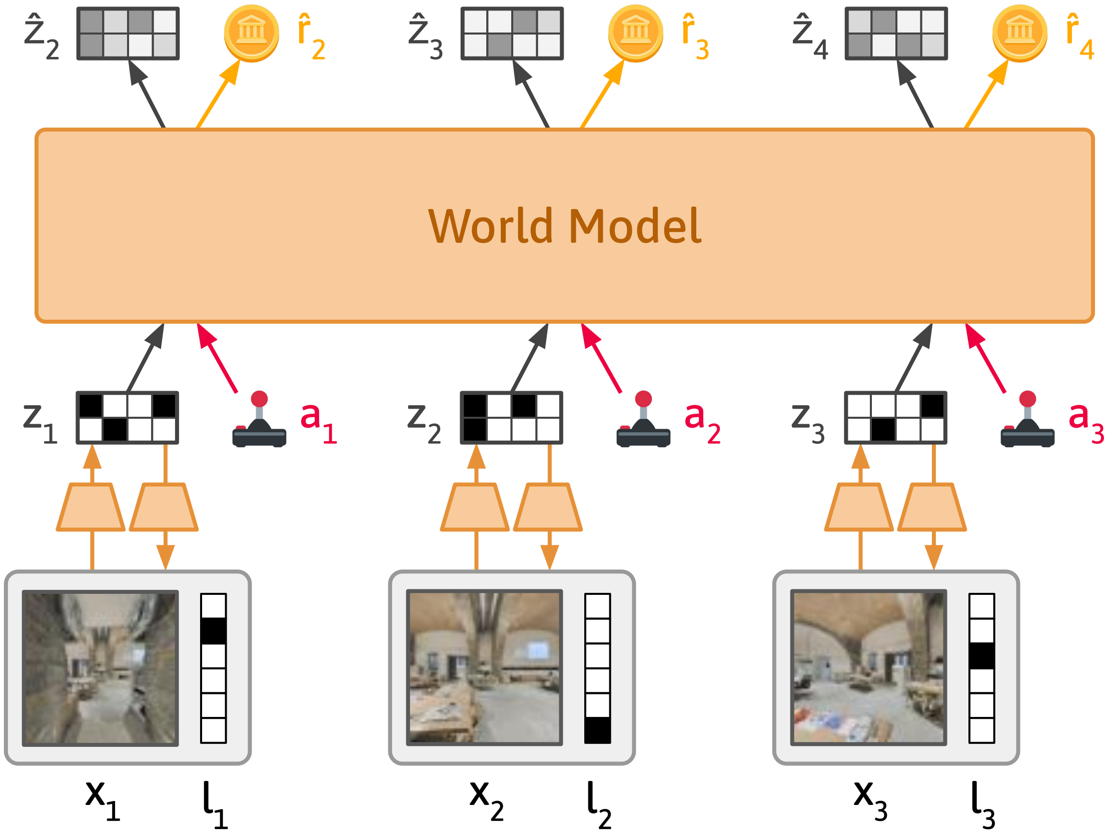
Dynalang



Dynalang



Dynalang



$\hat{z}_2$
$\hat{r}_2$
$\hat{z}_3$
$\hat{r}_3$
$\hat{z}_4$
$\hat{r}_4$
World Model
z_1
a_1
z_2
a_2
z_3
a_3
x_1
l_1
x_2
l_2
x_3
l_3

Dynalang

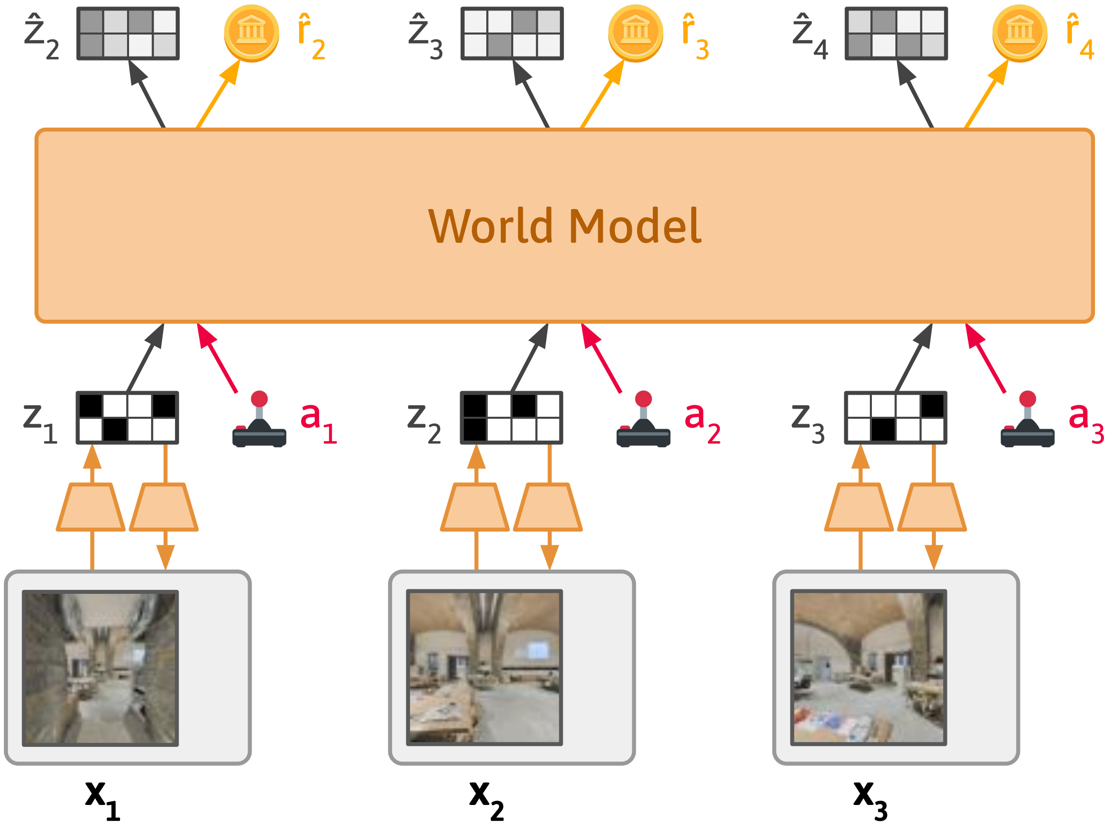
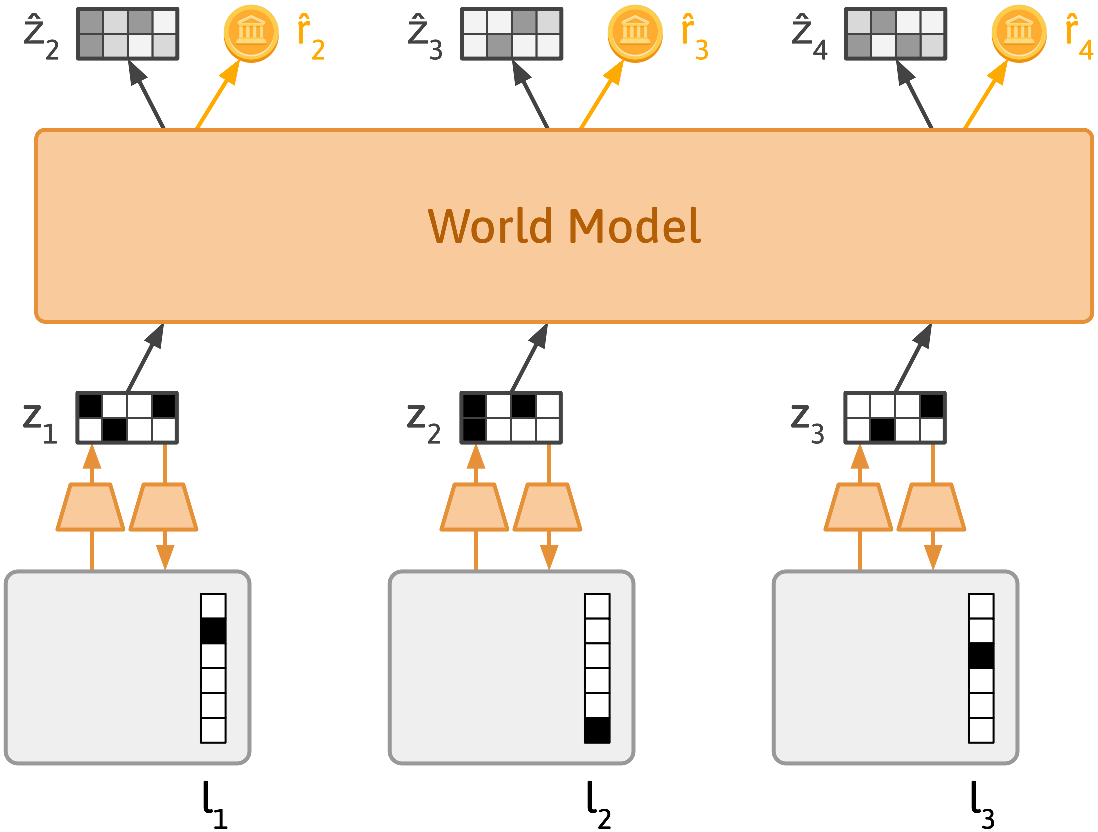


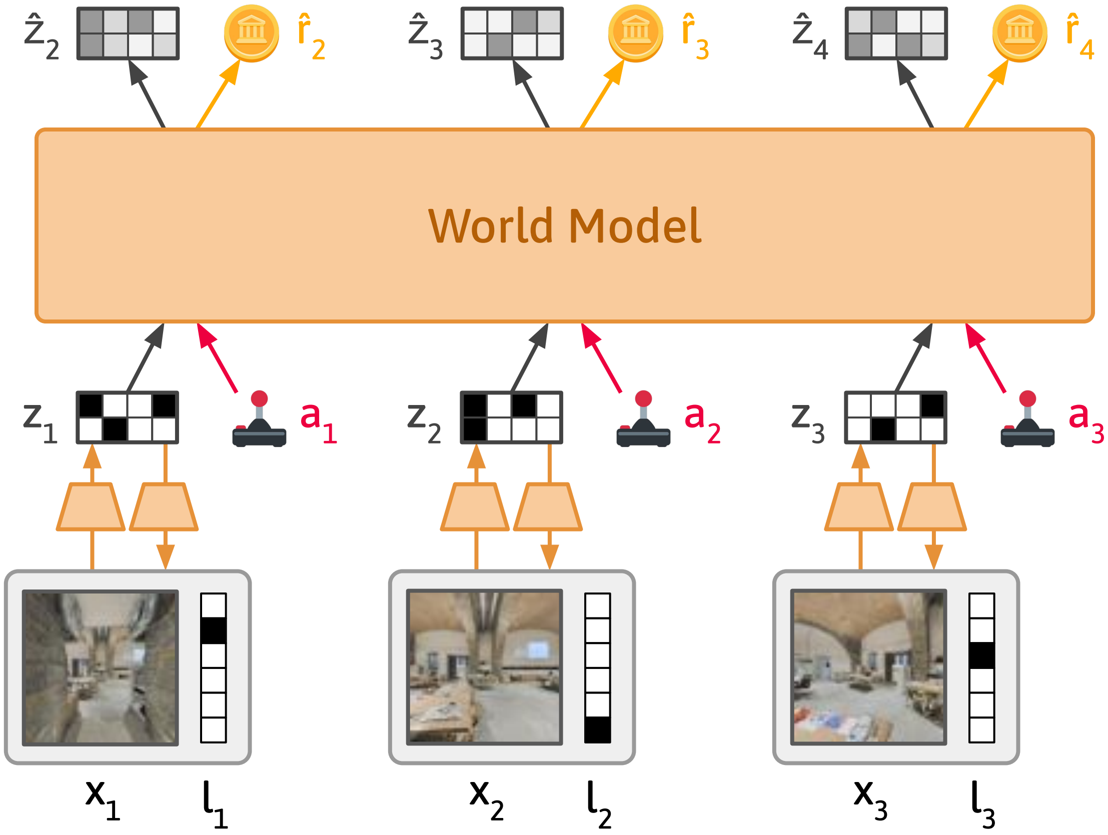
Image only: Video Prediction Model

Dynalang



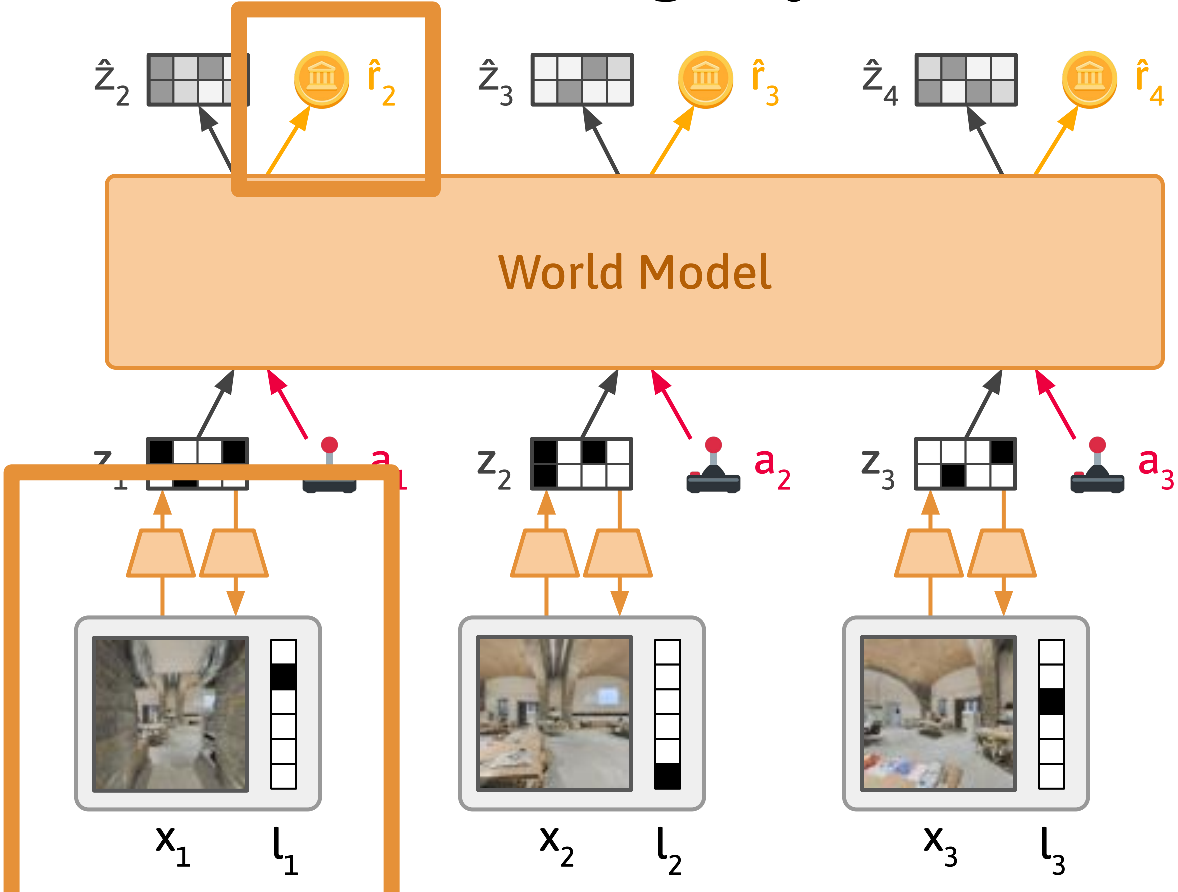
Text only:
(Latent-space)
Language Model

Dynalang



Model language tokens and images in a **joint latent representation that evolves over time.**

World Model Training Objective



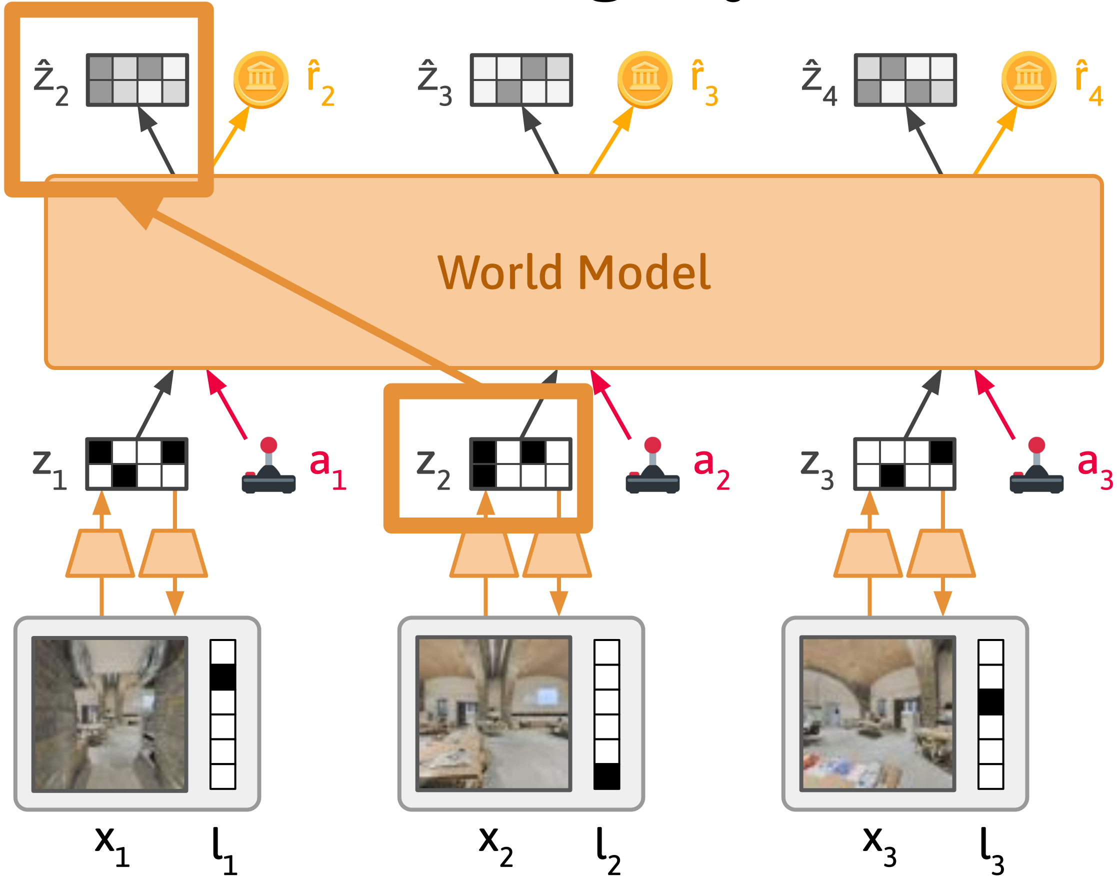
Reconstruction

Image

Text

Reward / Continue

World Model Training Objective



Reconstruction

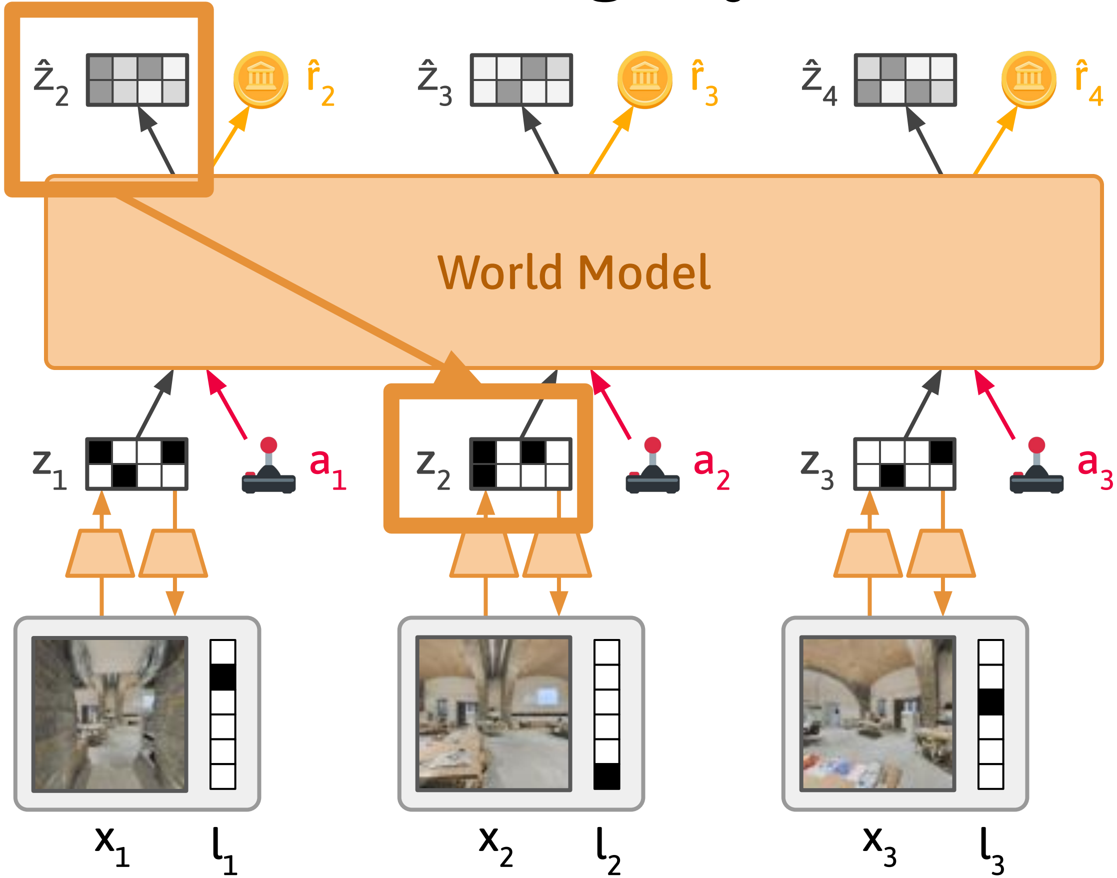
Image

Text

Reward / Continue

Representation Regularizer (L_{reg})

World Model Training Objective

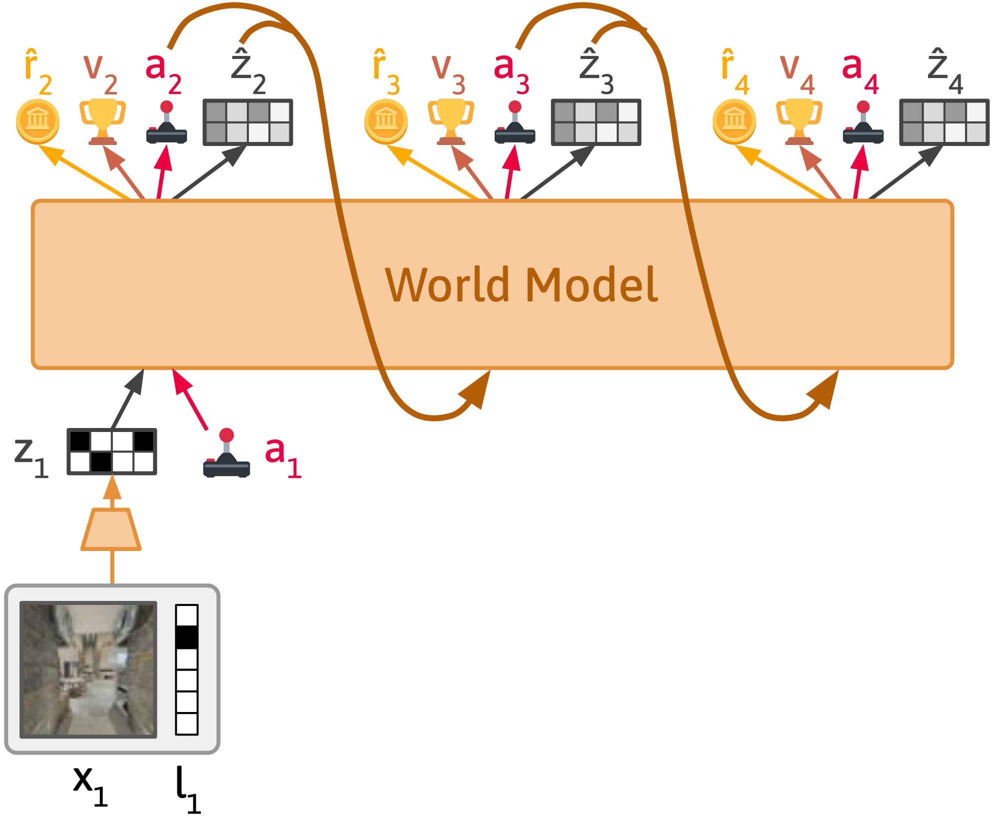


Reconstruction
- Image
- Text
- Reward / Continue

Representation Regularizer (L_{reg})

Future Prediction (L_{pred})

Learning to Act



Experiments

RL

Pretraining

Can we learn to use diverse kinds of language to better solve tasks, while maintaining instruction following abilities?

HomeGrid

Partially observed, multi-task environment with language hints

Reward
"Put the bottle in the recycling bin."

State
"The dishes are on the dining table."

Dynamics
"You need to press the pedal to open the trash can."

Corrections
"No, turn around."

Changing state
"I moved the dishes to the kitchen."
(coordination!)

HomeGrid

Partially observed, multi-task environment with language hints

Reward
"Put the bottle in the recycling bin."

State
"The dishes are on the dining table."

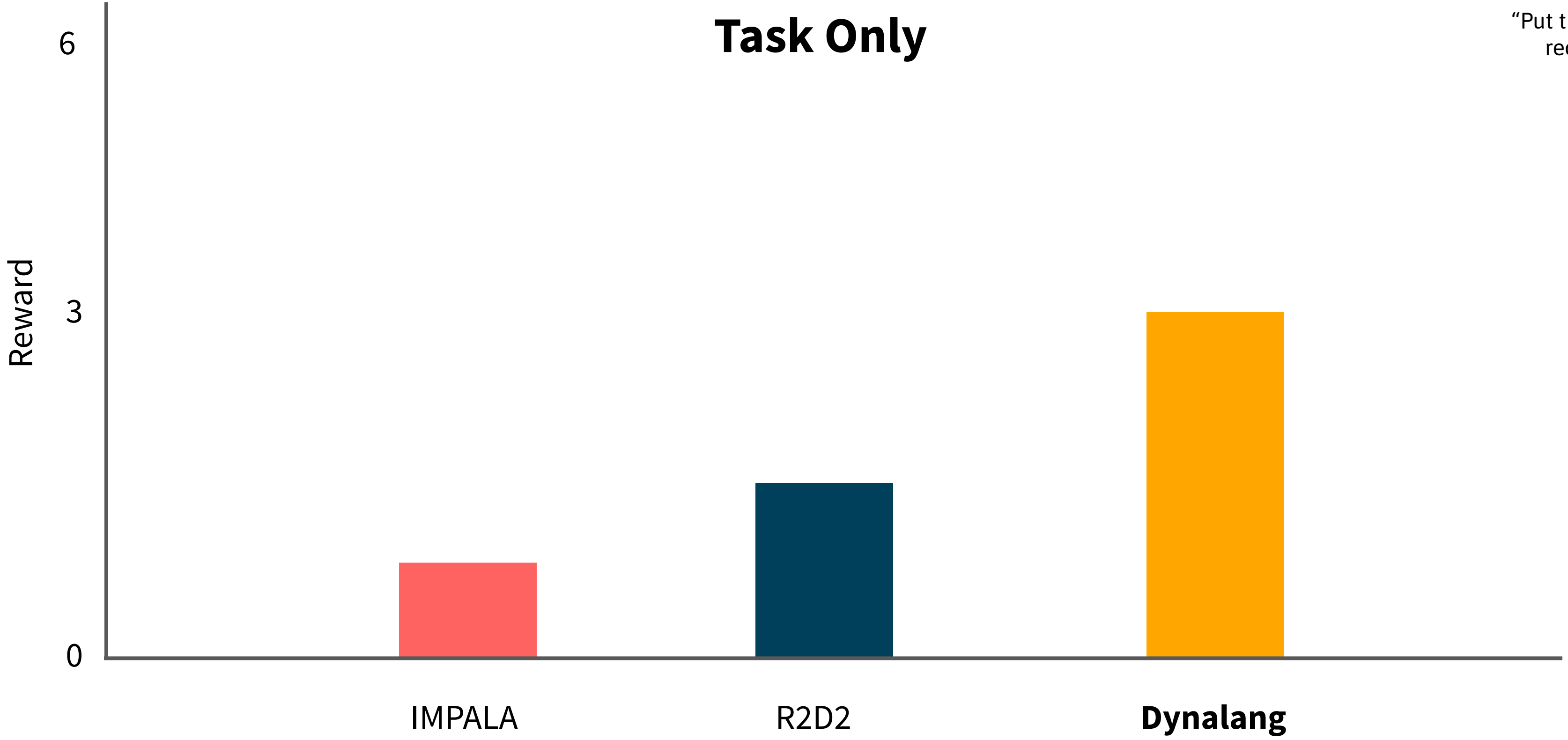
Dynamics
"You need to press the pedal to open the trash can."

Corrections
"No, turn around."

Changing state
"I moved the dishes to the kitchen."
(coordination!)

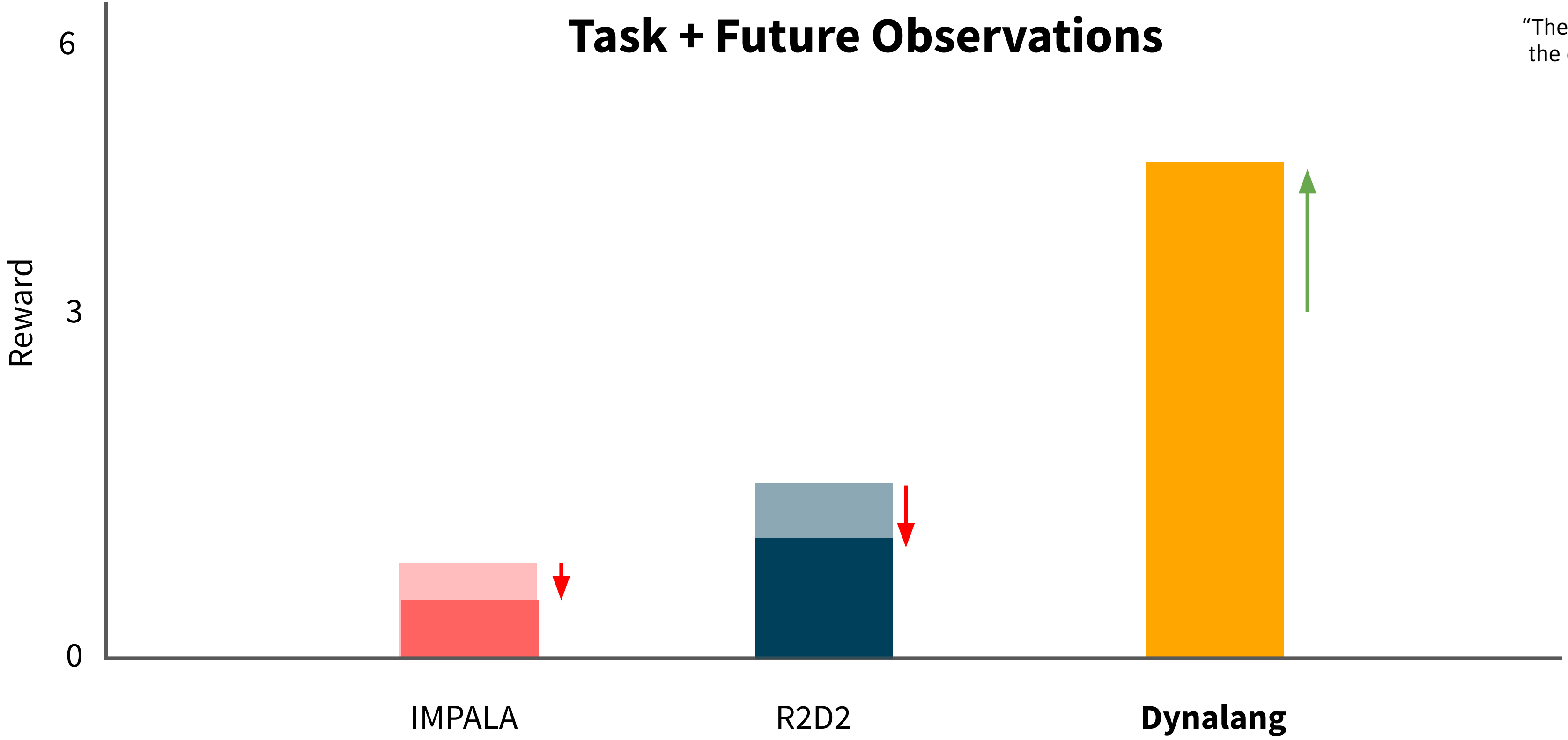
HomeGrid

Reward
“Put the bottle in the recycling bin.”



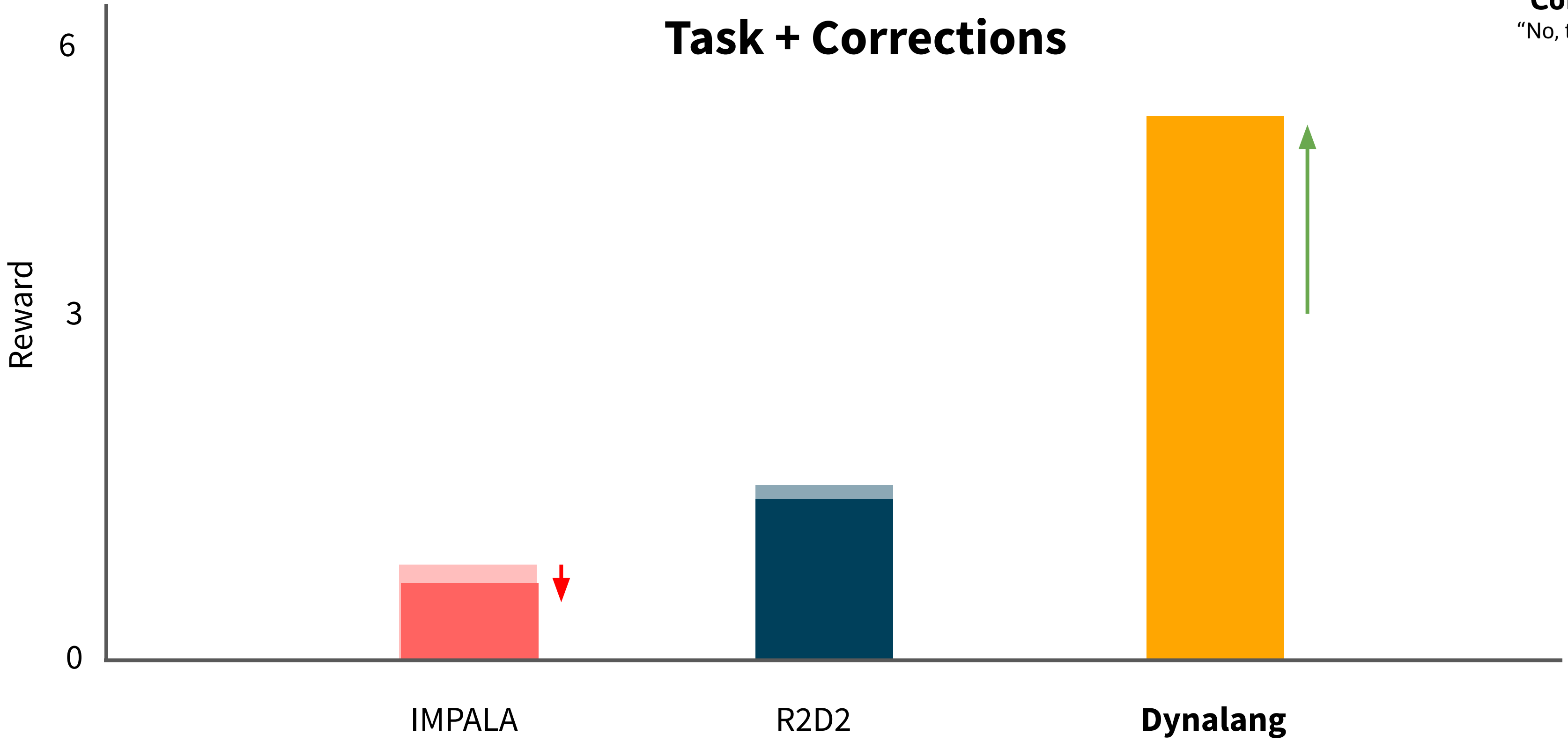
HomeGrid

State
"The dishes are on the dining table."



HomeGrid

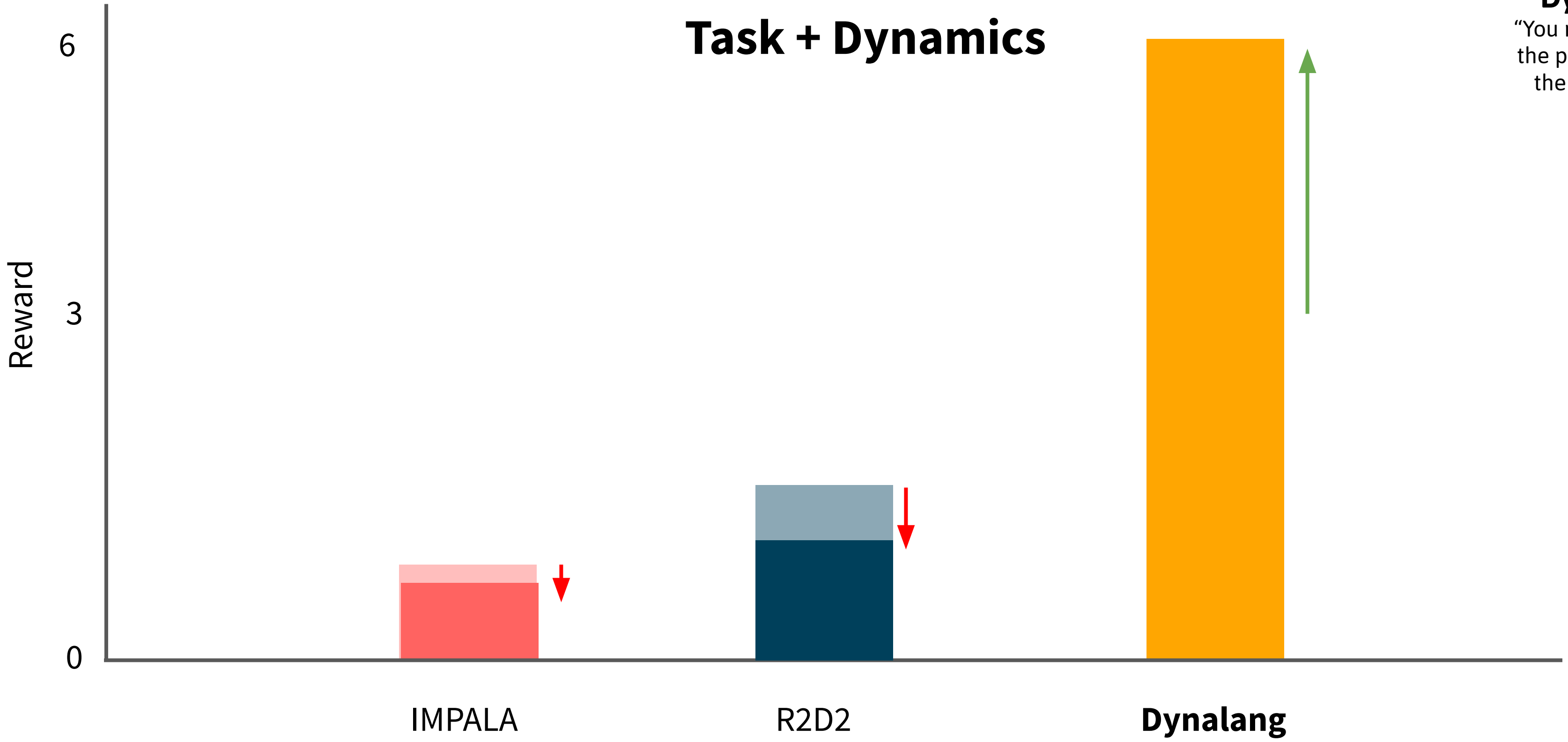
Corrections
"No, turn around."



Task + Corrections
6
3
0
Reward
IMPALA
R2D2
Dynalang

HomeGrid

Dynamics
"You need to press the pedal to open the trash can."



Decoding from Model Representations

Context

Video and text inputs

the **bottle** is in the living room

get the **bottle**

the **plates** are in the

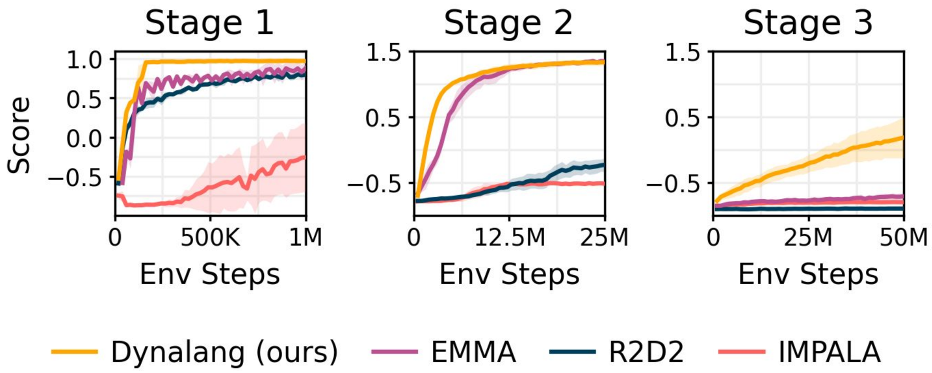
Dynalang Model Rollouts

Video prediction

Messenger Multi-hop reasoning with manuals

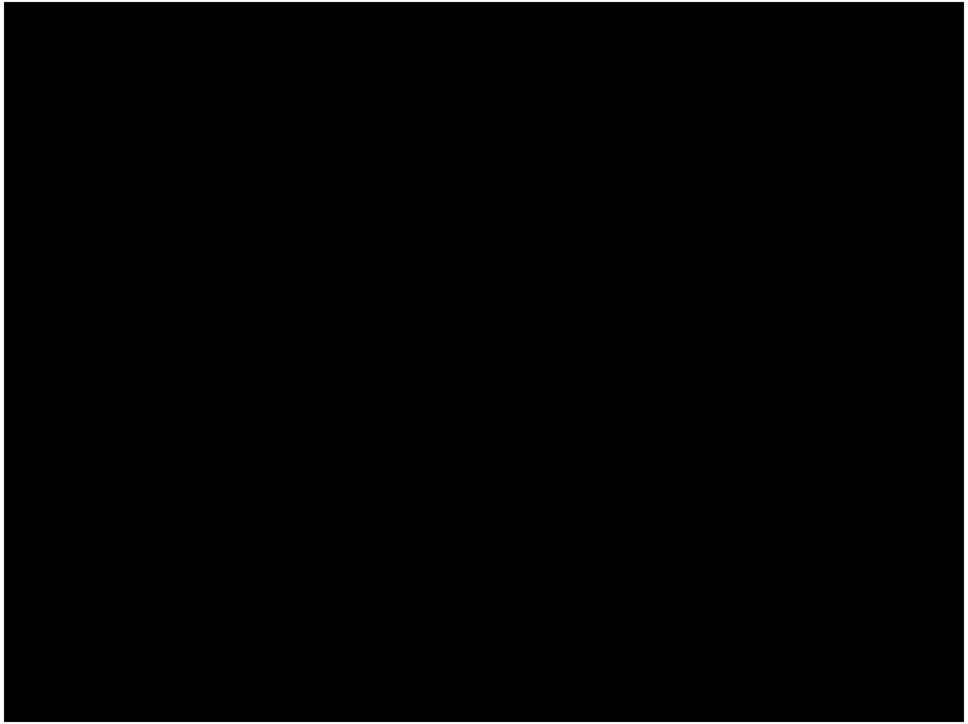
GAME 1 MANUAL

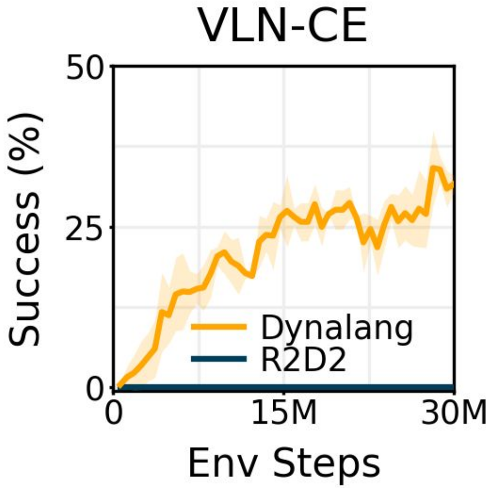
1. at a particular locale, there exists a motionless mongrel that is a formidable adversary.
2. the top-secret paperwork is in the crook's possession, and he's heading closer and closer to where you are.
3. the crucial target is held by the wizard and the wizard is fleeing from you.
4. the mugger rushing 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.*



Vision-Language Navigation

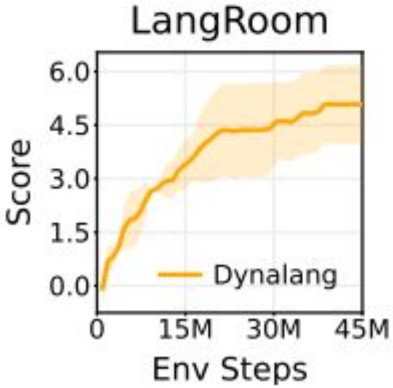
Photorealistic instruction following





LangRoom

Experience-informed language generation

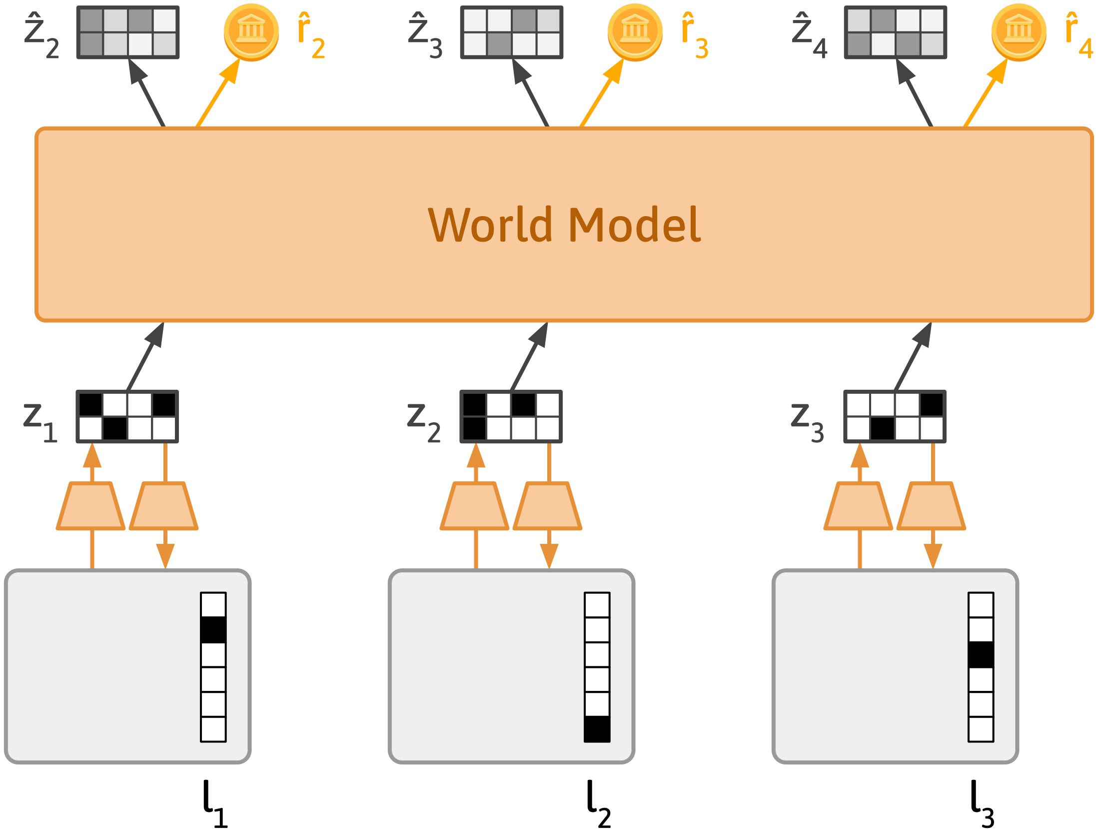


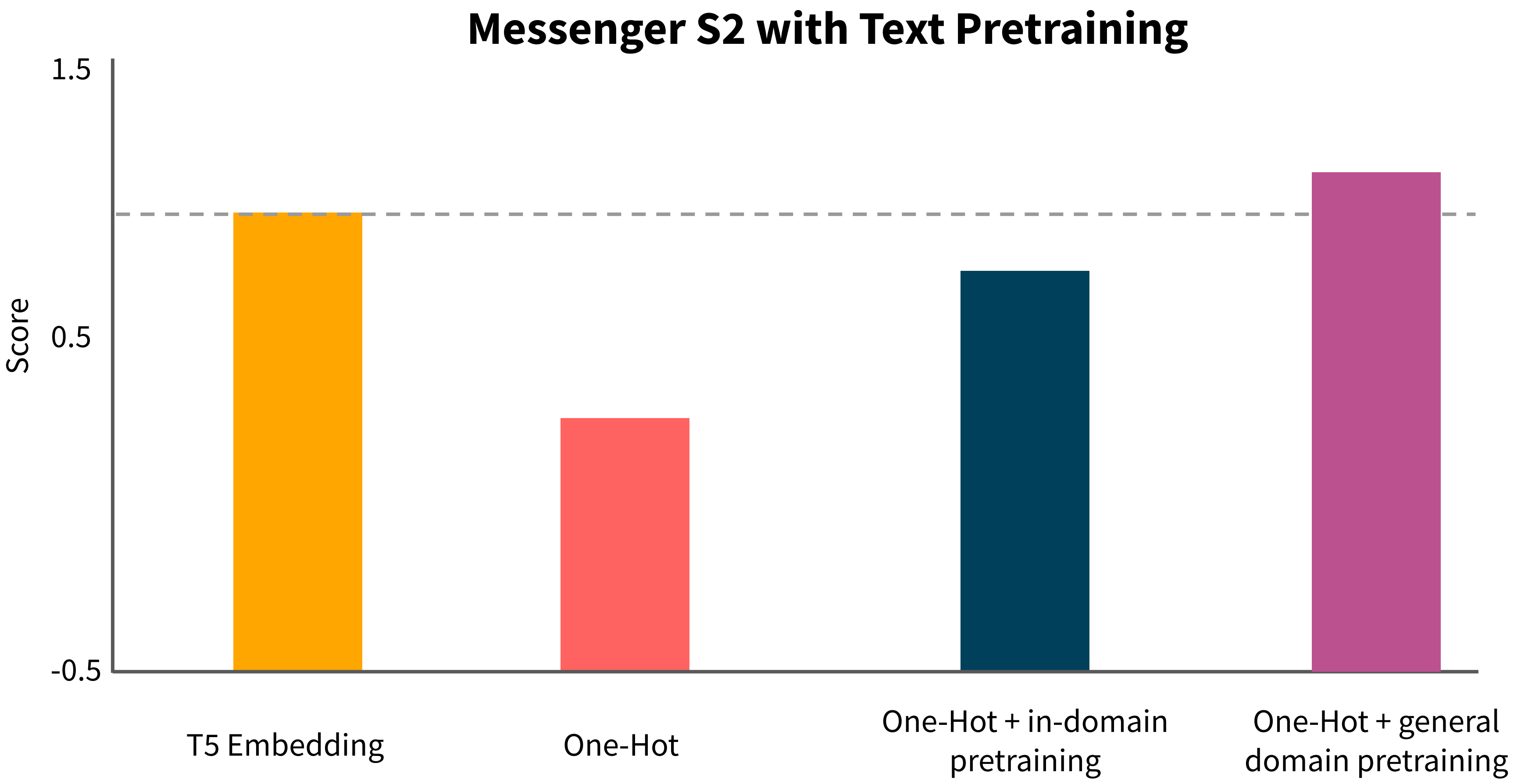
So far: we can improve online with language and vision inputs

…but we generally don't want to learn knowledge from scratch.

Is this approach compatible with large-scale pretraining?

Scaling Up with Pretraining




Messenger S2 with Text Pretraining
1.5
0.5
-0.5
Score
T5 Embedding
One-Hot
One-Hot + in-domain pretraining
One-Hot + general domain pretraining

Summary

To solve tasks in the real world, interactive agents will need to understand how language relates to the world around them.

Language as an observation that helps us predict the future suggests a unified learning objective for language, vision, and actions: **self-supervised prediction inside a multimodal world model.**

This enables Dynalang to learn both *offline* from large-scale action-free data and *online* from experience.

Thank You!

Paper + Code @ dynalang.github.io

@realJessyLin jessy_lin@berkeley.edu