Digital Human

ZHENG Zishuo

2023.08.08

Highlights

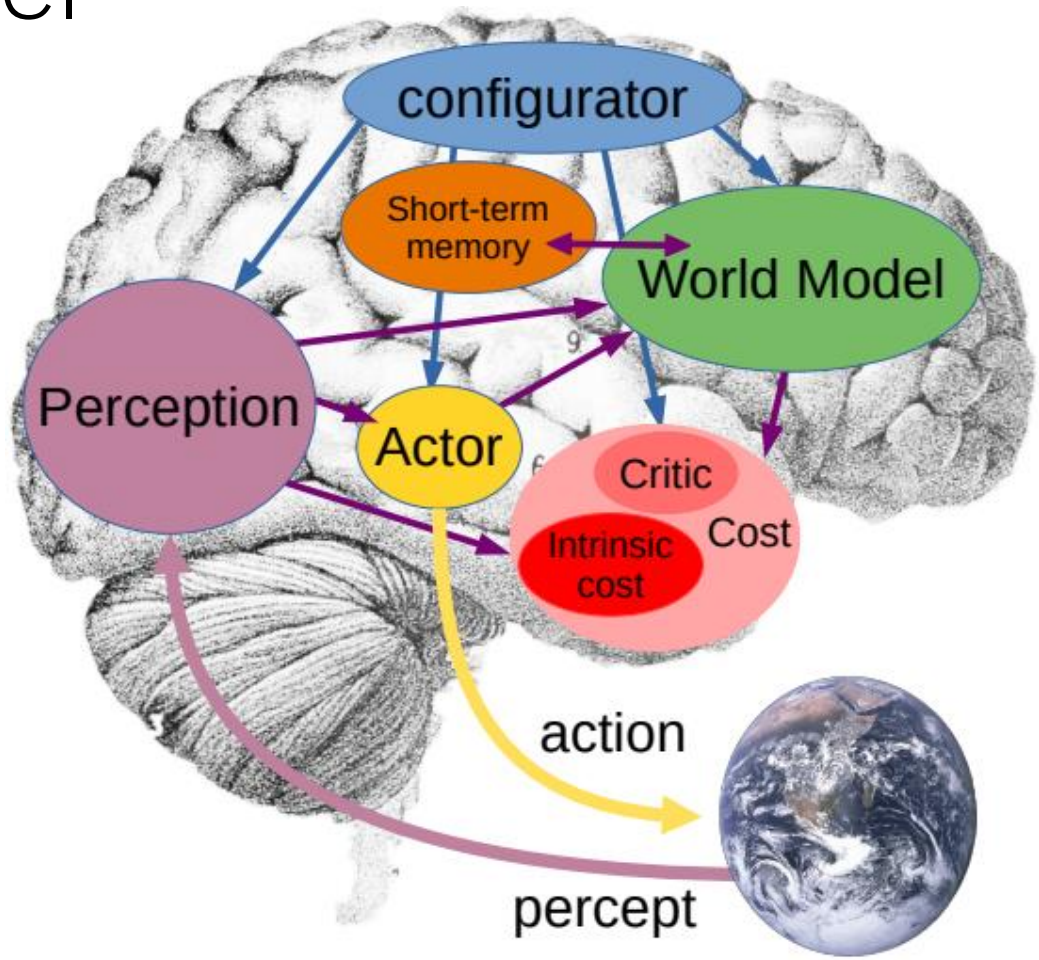
- The agent can grow up in the environment without any additional information except active RGBD perception and proprioception.
- An autonomous universal agent framework for any robot & any environment.

Setup: Env & Agent

https://github.com/threedworld-mit/tdw

https://github.com/alters-mit/magnebot

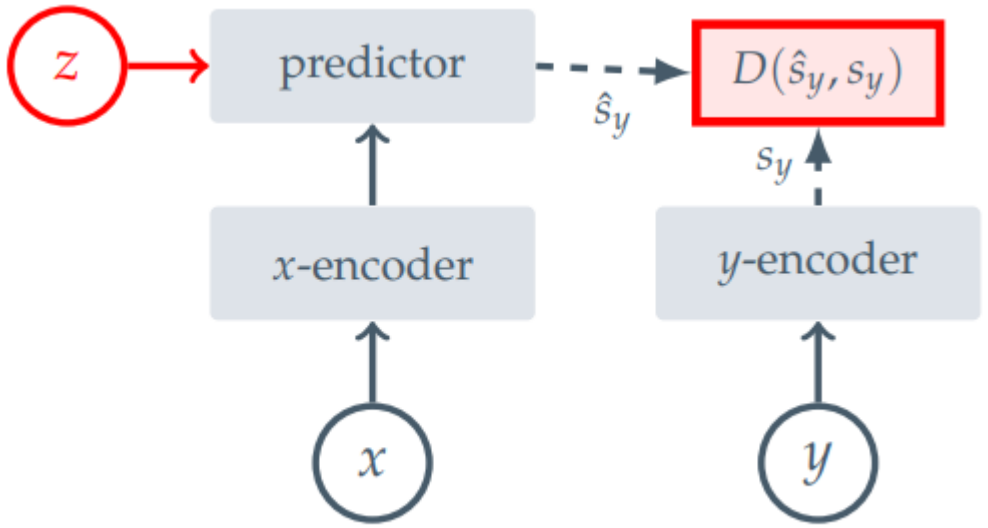
World Model



Yann LeCun, 2022

Stage1: Self-Control (Baby Room)

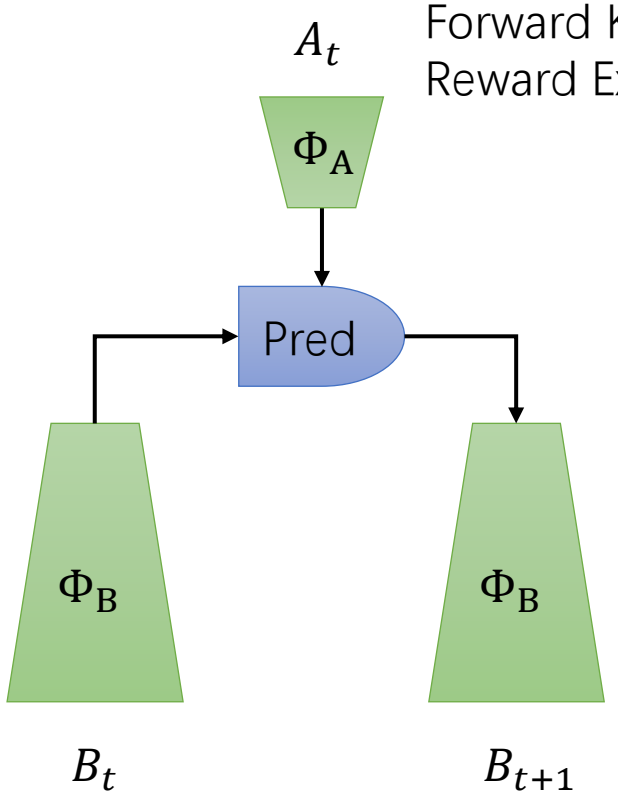
Do body states and action tokens need information selection?



(c) **Joint-Embedding Predictive Architecture**

JEPA for SSL

M Assran et al, CVPR, 2023



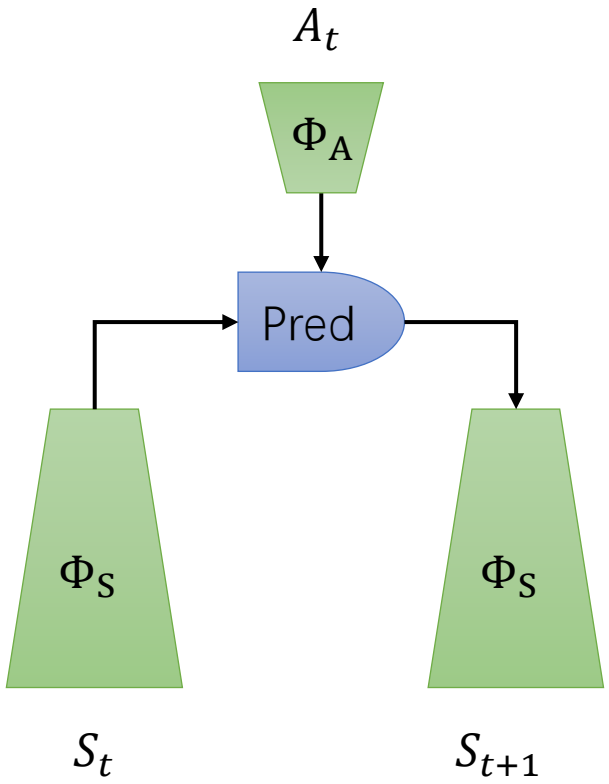
Forward Kinematics
Reward Exception Error

Inner System:
The body is deterministic

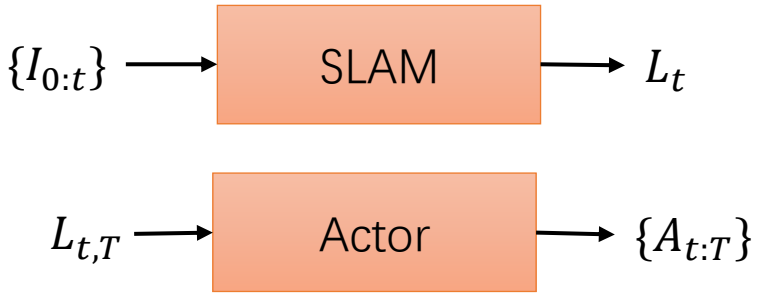
A: actions {joint motions}
B: body states {component coordinates}
Φ: embedding network

T-JEPA
Temporal Joint-Embedding Predictive Architecture

Stage 2: Locomotion & SLAM (Clean Room)



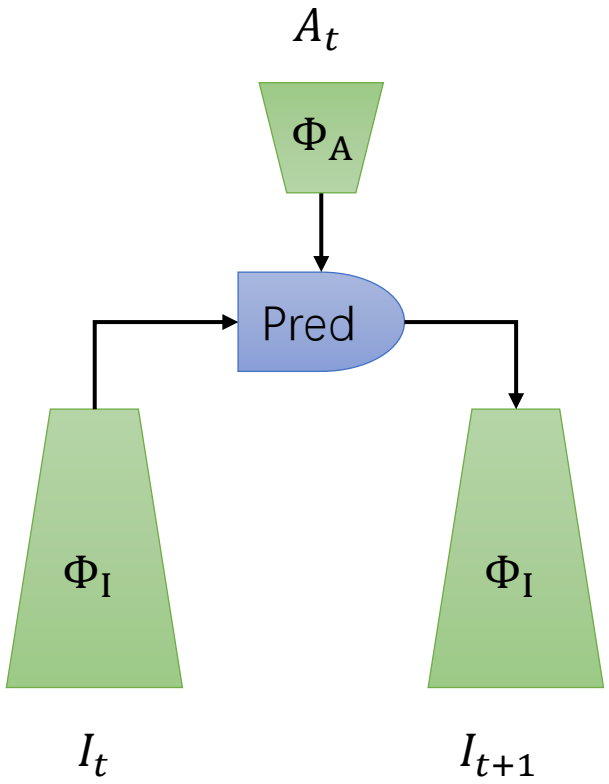
1. Locomotion:
   a) Safety: RGBD cloud points and self-body should not collide.
   b) The final coordinate as a reward, search for action rollout
2. SLAM: explore unknown areas by designing a reward function.
   a) Distance can be read from the SLAM model.
   b) Build coordinate system.
3. **Faults: Locomotion is binding with the environment. Cloud point-based scene representation as input may solve this issue.**
4. **SLAM and location coordinates can also use JEPA with I_t location proxy. The target is the image patch.**



Inner System Changes
Egocentric States

I: partial observation, RGBD
L: location coordinate
$S_t = [B_t; L_t]$
Φ_S can be finetuned from Φ_B

Stage 3: Manipulation & Prediction (Object Room)



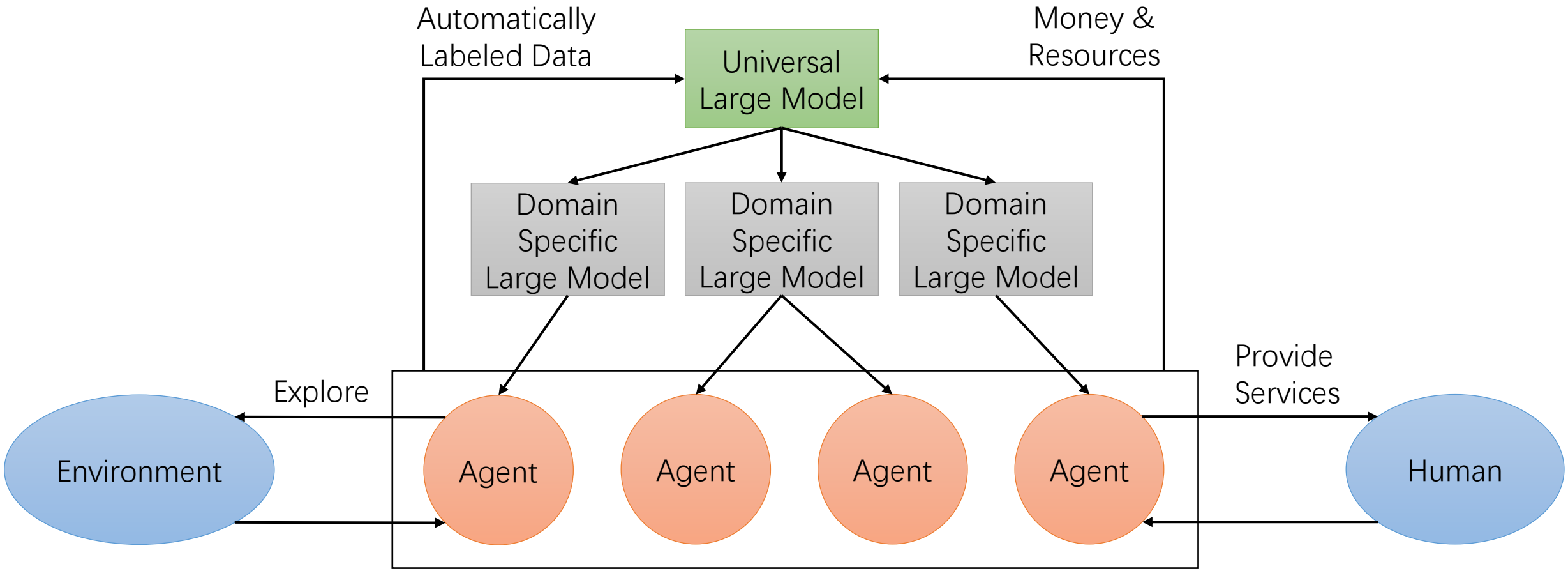
1. Random Select target, then move to the target and try to manipulate.
2. Inner encoder Φ and outer encoder should be separate. From egocentric to allocentric, states have increasing levels of instability and variance.
3. Manipulation (Only Rigid Body Now)
   a) RGBD cloud points and self-body should collide.
   b) Latent segmentation and 3D model should emerge by itself. (like optical flow. The seg map is an interface for humans. It has no meaning for the agent.)
   c) Share the same actor system with locomotion.
   d) Actor and predictor are trained iteratively.

(easy-to-predict principle, minimum energy)

Inner System Changes
Allocentric Object States

Stage 4: Honeycomb Data Loop

How to build a large model automatically?



Future Work

- In LeCun's World Model
  - MultiModality
  - Task Switch & Configuration Module
  - Hierarchical Abstraction / Planning
- Larger Scale and More Complex Environment
- Generative Model & Train by Imaging
- Combine Bottom-Up Hierarchical Abstraction with Top-Down Configuration