CSCI1470

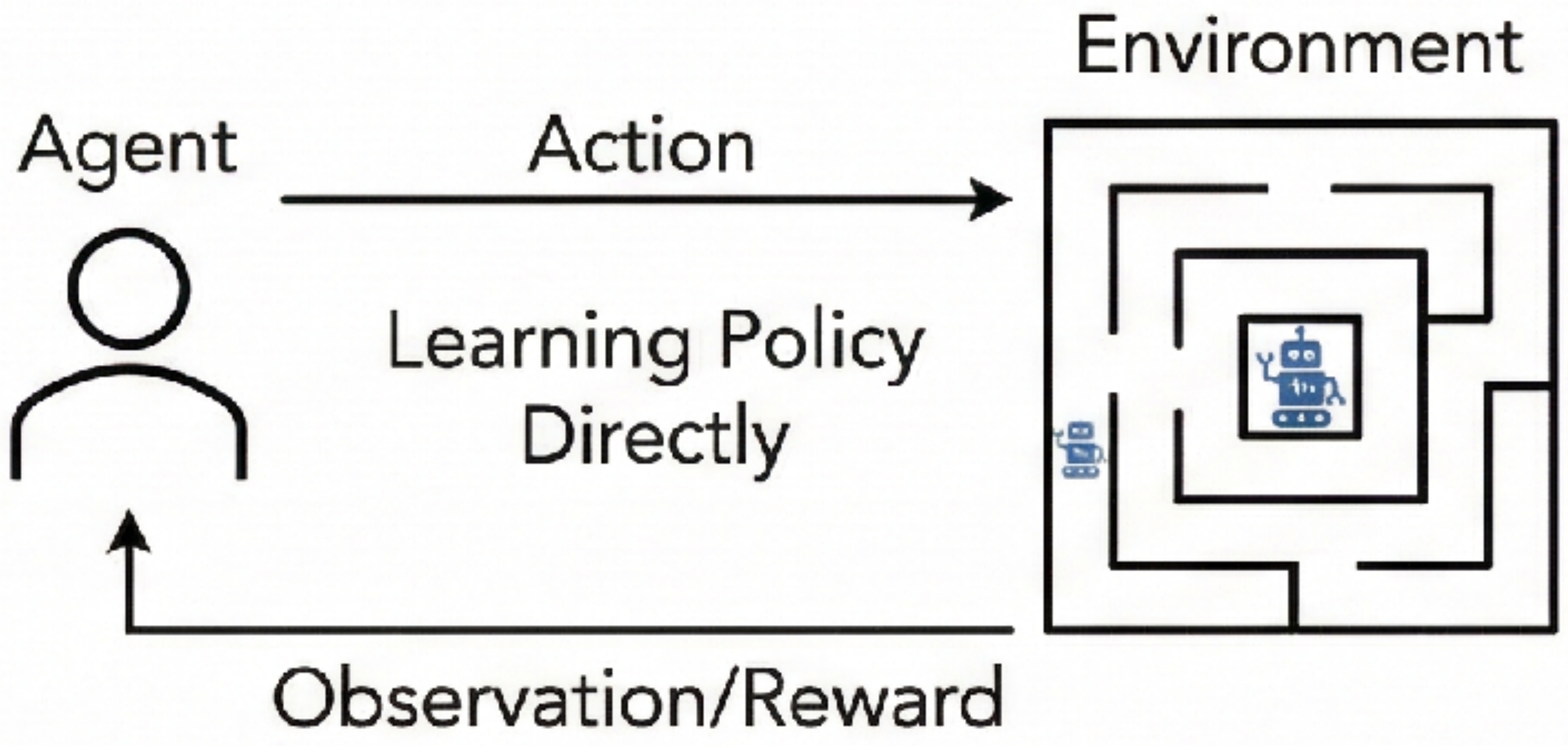
Deep Learning

Randall Balestriero

Recap

MODEL-BASED REINFORCEMENT LEARNING: THE CONCEPT

A: MODEL-FREE RL



Environment

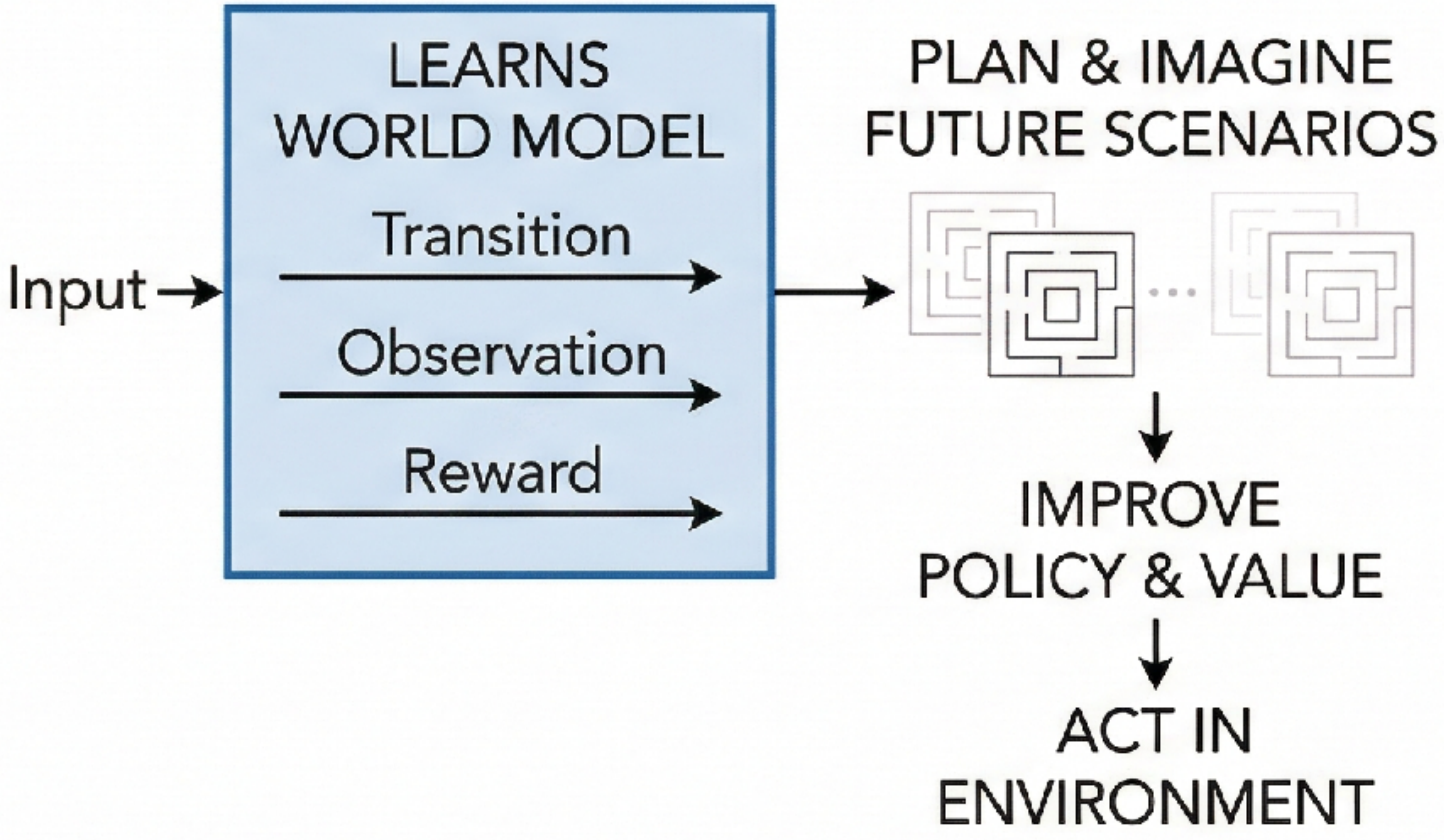
Agent

Action

Learning Policy Directly

Observation/Reward

B: MODEL-BASED RL



LEARNS WORLD MODEL

Transition

Observation

Reward

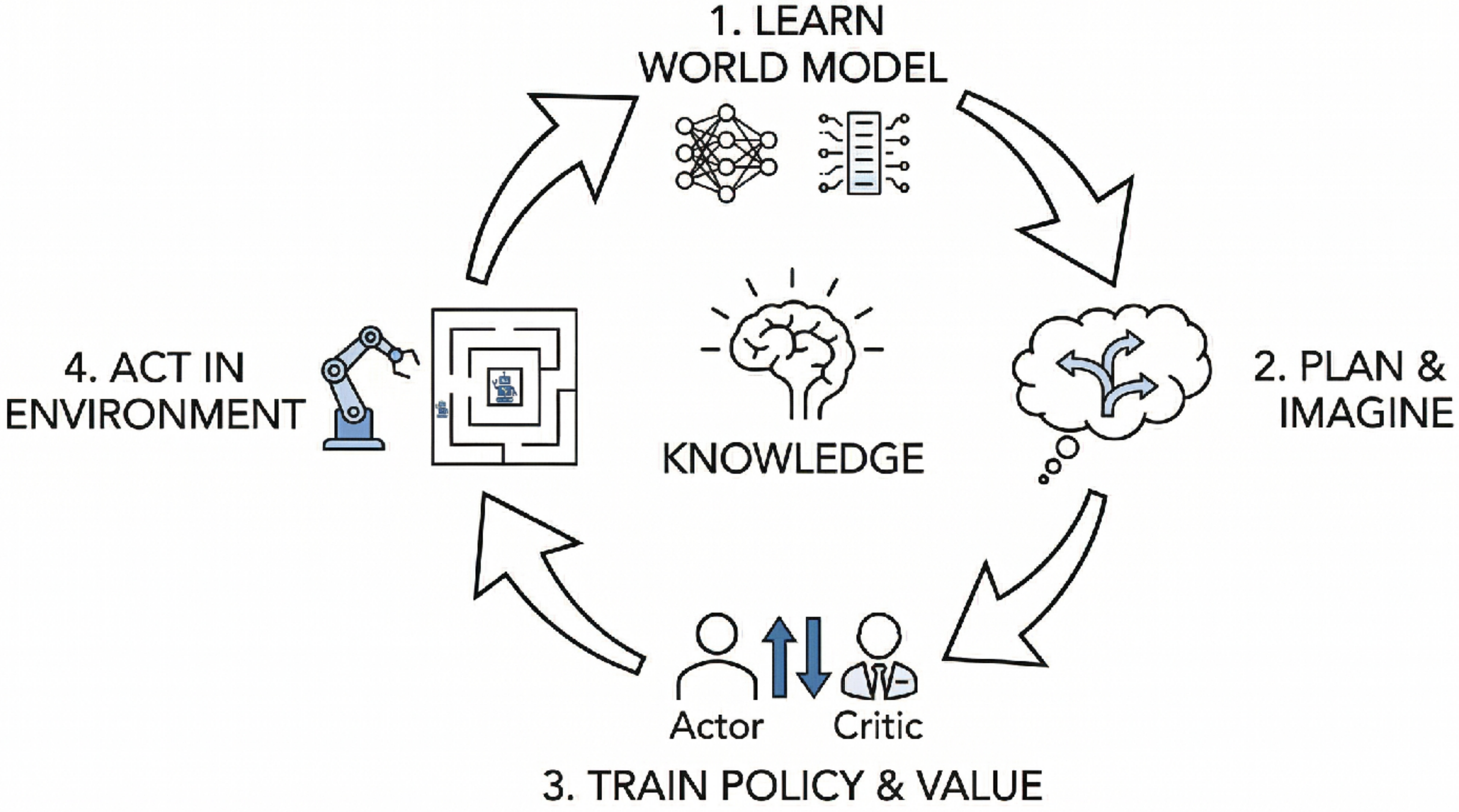
Input

PLAN & IMAGINE FUTURE SCENARIOS

IMPROVE POLICY & VALUE

ACT IN ENVIRONMENT

THE MODEL-BASED RL CYCLE



1. LEARN WORLD MODEL

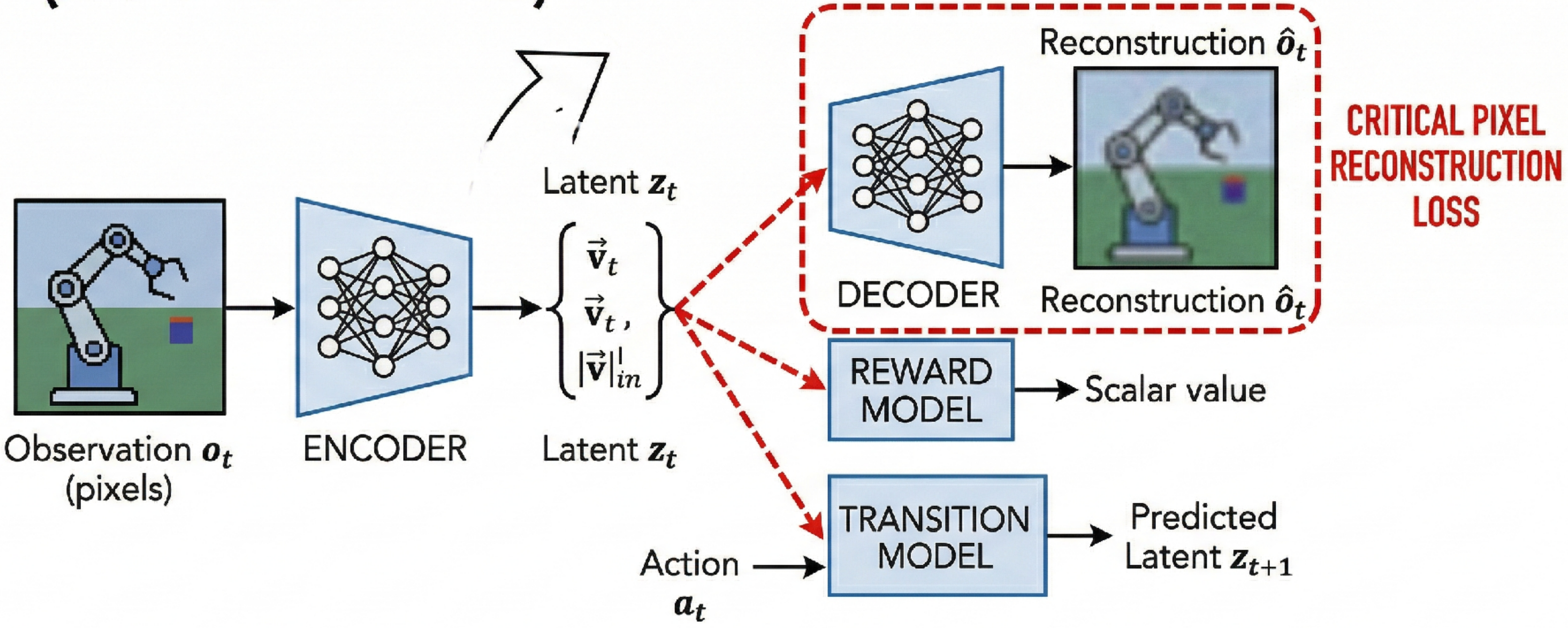
2. PLAN & IMAGINE

3. TRAIN POLICY & VALUE

4. ACT IN ENVIRONMENT

KNOWLEDGE

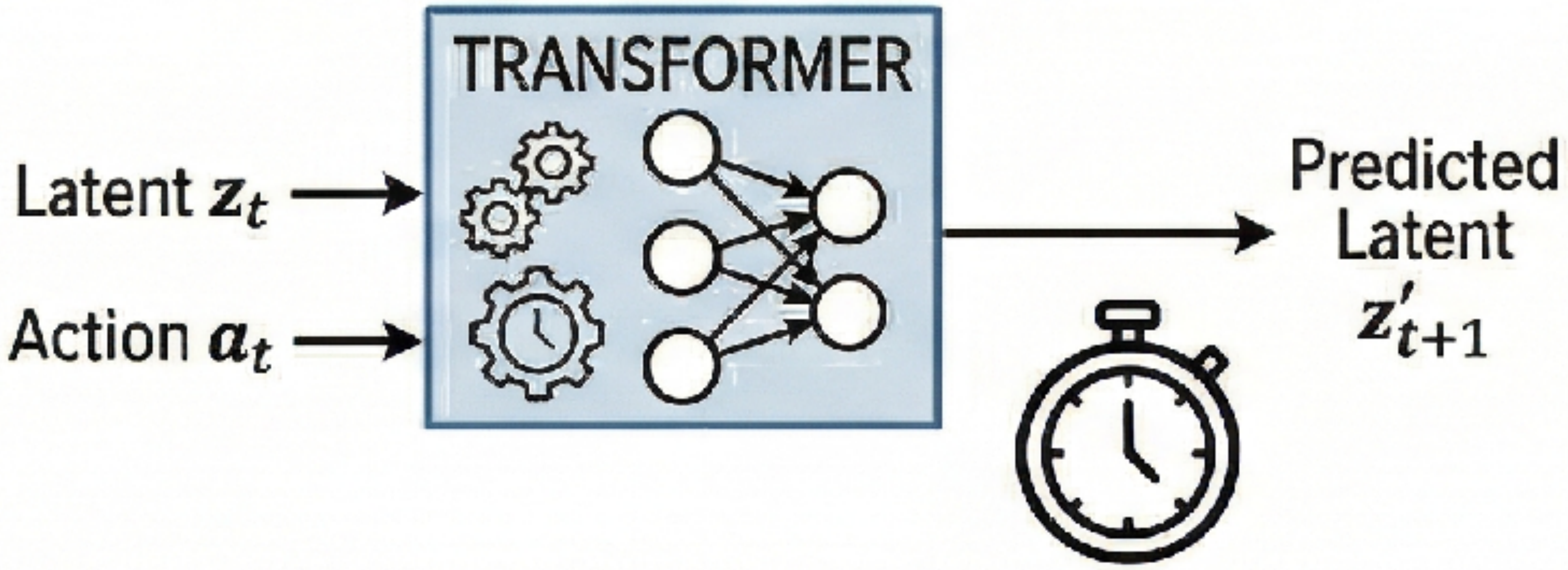
EXAMPLE: PIXEL-SPACE WORLD MODEL (DREAMER-v3)



Latent $\boldsymbol{z}_t$
Observation $\boldsymbol{o}_t$ (pixels)
ENCODER
$\vec{\mathbf{v}}_t$
$\vec{\mathbf{v}}_t$,
$|\vec{\mathbf{v}}|_{in}^{|}$
Latent $\boldsymbol{z}_t$
Reconstruction $\hat{\boldsymbol{o}}_t$
DECODER
Reconstruction $\hat{\boldsymbol{o}}_t$
CRITICAL PIXEL RECONSTRUCTION LOSS
REWARD MODEL
Scalar value
Action $\boldsymbol{a}_t$
TRANSITION MODEL
Predicted Latent $\boldsymbol{z}_{t+1}$

DreamerV4!

1. TRANSFORMER-BASED TRANSITION MODEL



TRANSFORMER

Latent z_t →

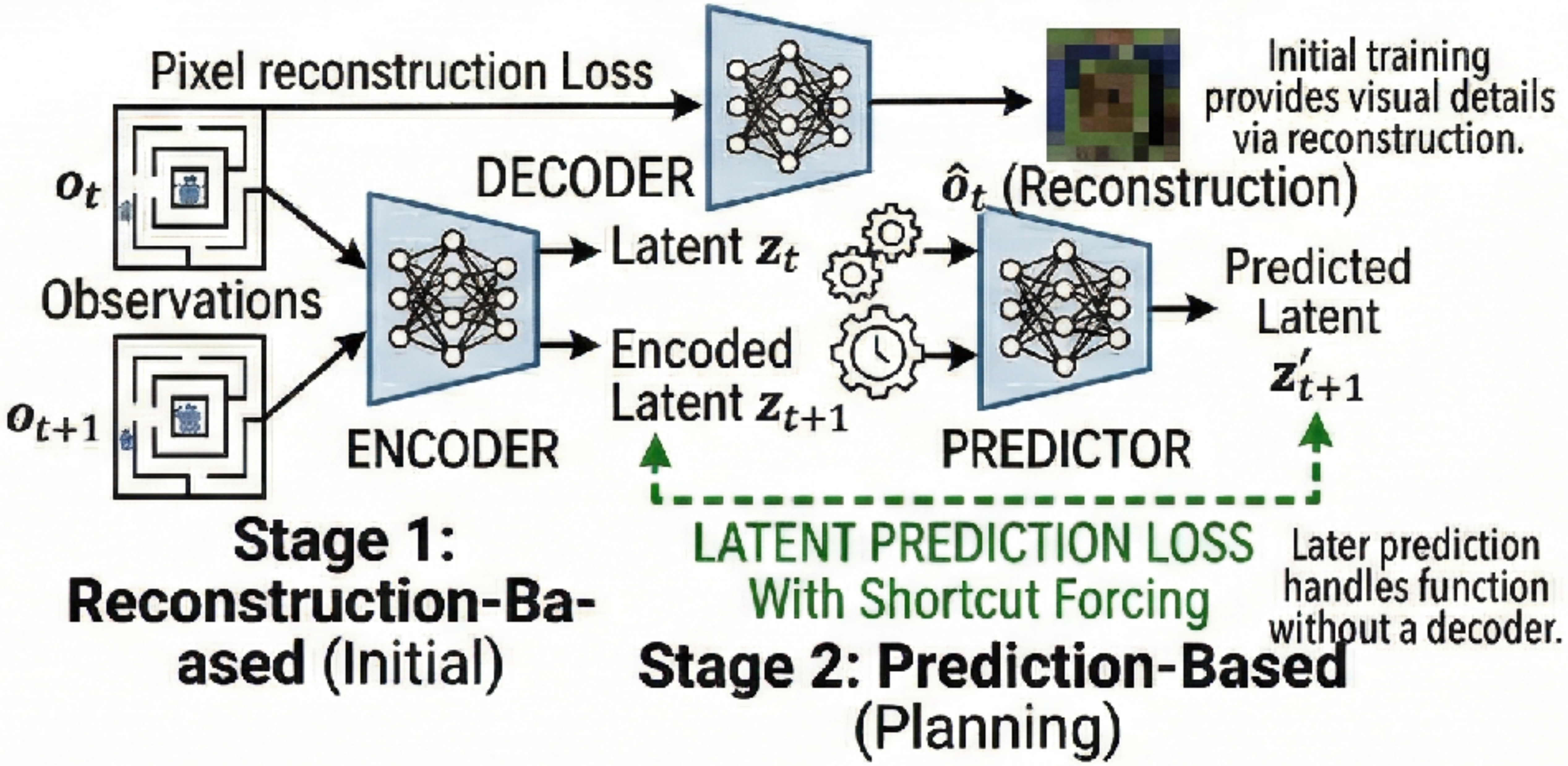
Action a_t →

Predicted Latent z'_{t+1}

*Models longer horizons.

*Scalable and computationally efficient.

2. HYBRID LATENT SPACE LEARNING



Pixel reconstruction Loss

DECODER

Initial training provides visual details via reconstruction.

$\hat{o}_t$ (Reconstruction)

o_t

Observations

o_{t+1}

ENCODER

Latent z_t

Encoded Latent z_{t+1}

PREDICTOR

Predicted Latent z'_{t+1}

LATENT PREDICTION LOSS With Shortcut Forcing

Later prediction handles function without a decoder.

Stage 1: Reconstruction-Based (Initial)

Stage 2: Prediction-Based (Planning)

*Initial training captures visual detail via pixel reconstruction.

*Later training predicts functional structure without a decoder.

TAKEAWAY: Data-Efficient & Scalable Real-Time Performance and Control from Pixels.

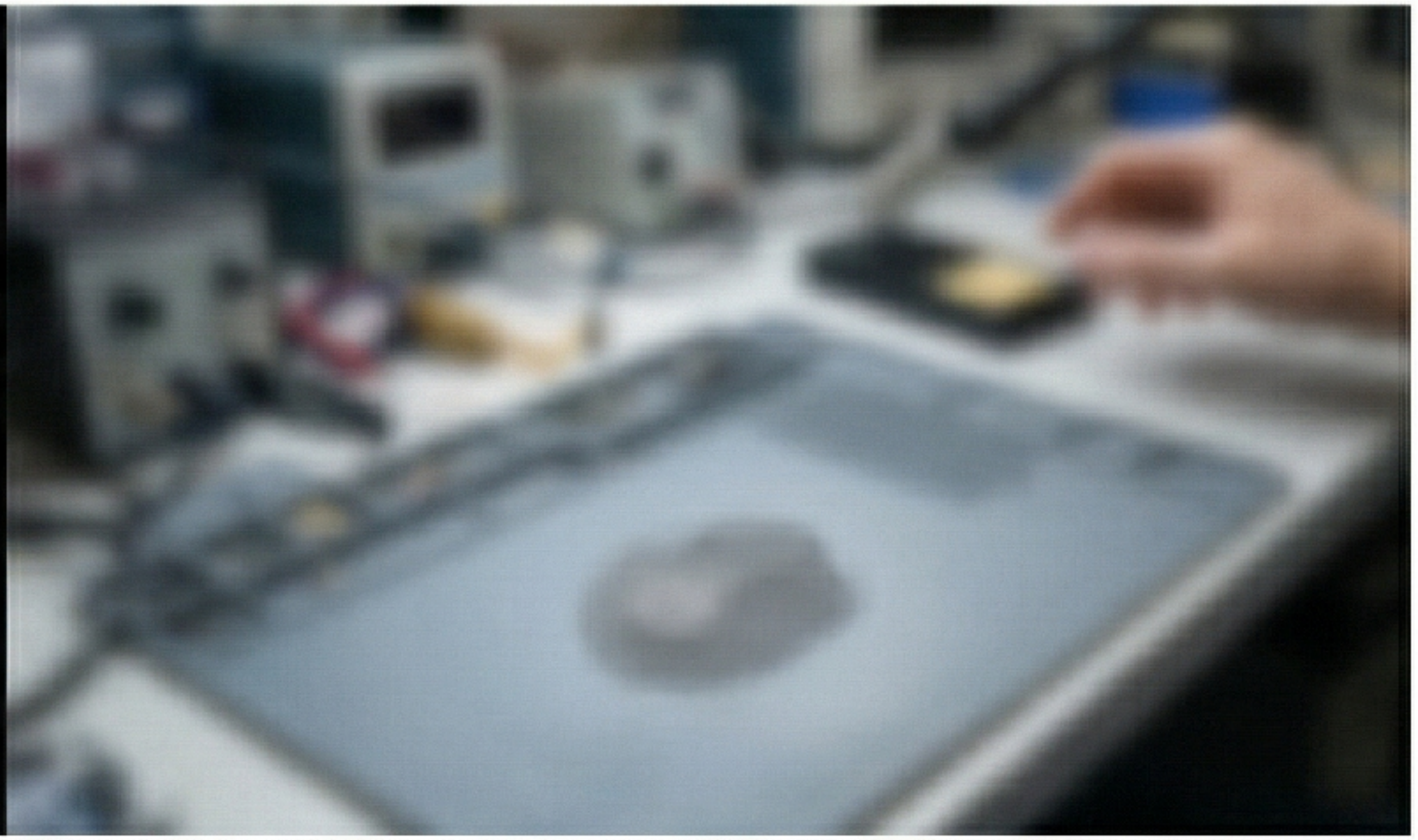
- Scales to large unlabeled datasets.

Fast real-time
(20 FPS+)

Offline ur-picking context
(Minecraft, from <IMAGE 6)

FOOD FOR THOUGHT

THE REALITY (o_t)



DECODED IMAGINATION ($\hat{o}_t$)

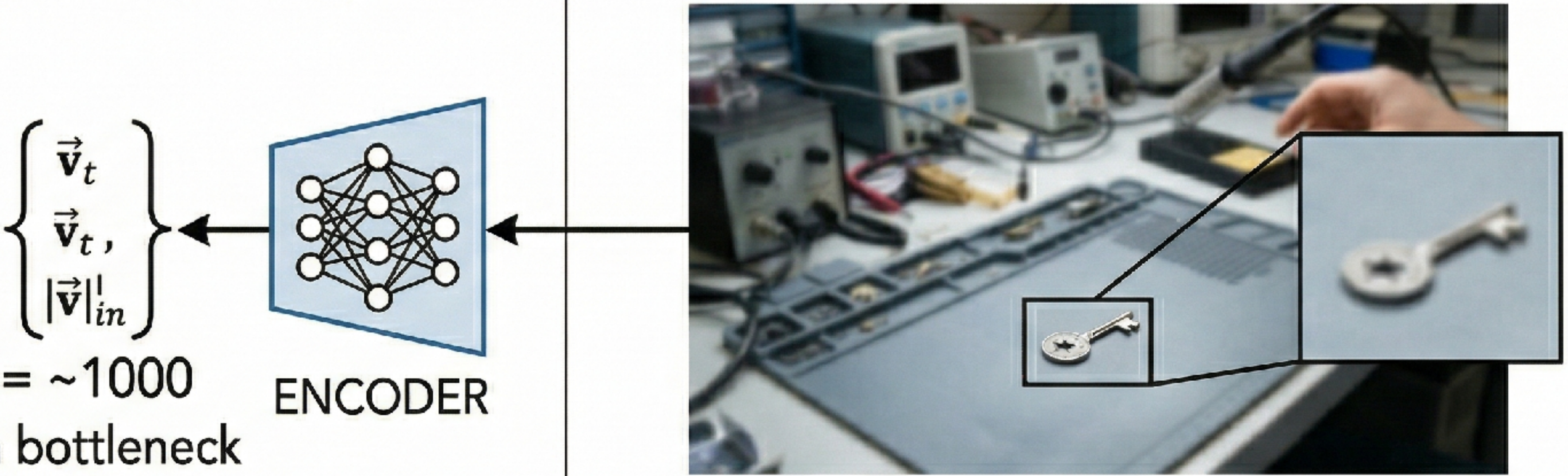
IS RECONSTRUCTING EVERY PIXEL A GOOD USE OF COMPUTATION?

Wasteful? Irrelevant details? Wrong metric?

WHY PIXEL RECONSTRUCTION IS HARD: THE DIMENSIONALITY CURSE

A: COMPACT LATENT (z_t)

B: HIGH-RES OBSERVATION (o_t)



d = ~1000

Information bottleneck

Easy to Plan

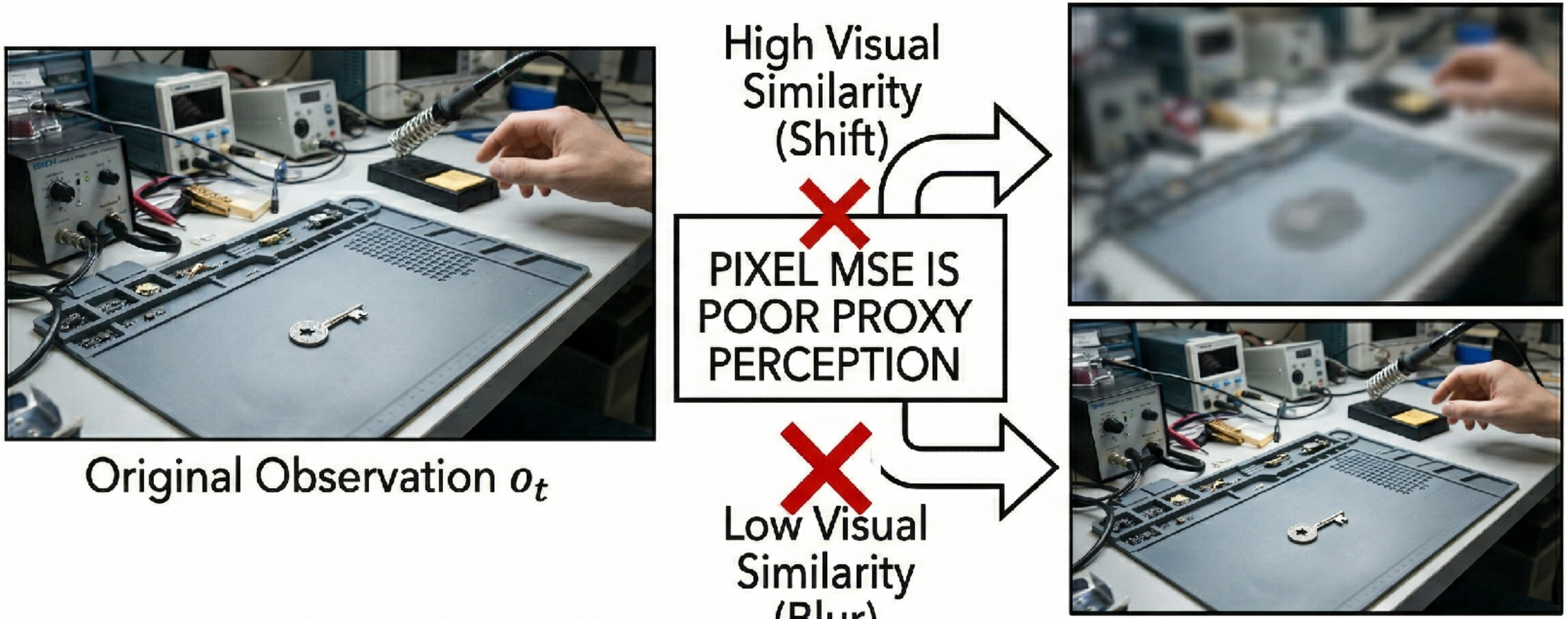
d = 640x360x3 = ~690,000

High redundancy, complex details

✗

Wasteful to Imagine

THE CORE CHALLENGE: COMPARING IMAGE OBSERVATIONS



Reconstruction $\hat{o}_t$ (MSE ≈ 50)

High Visual Similarity (Shift)

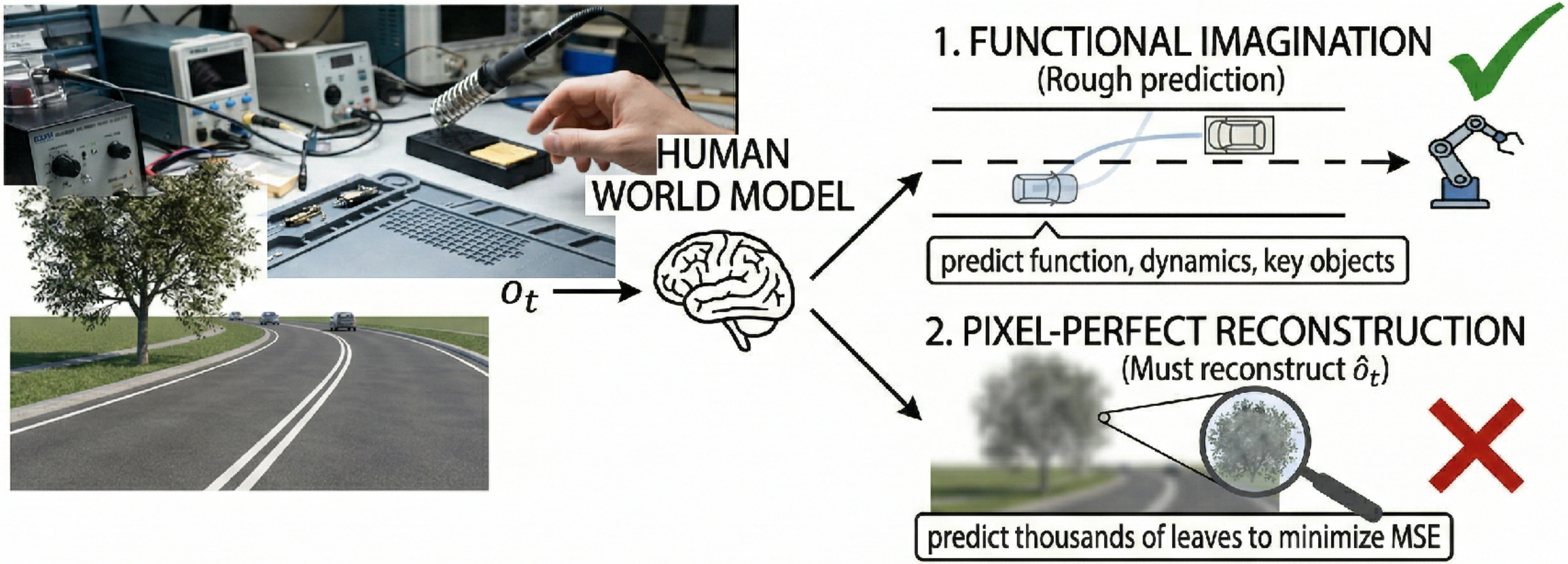
PIXEL MSE IS POOR PROXY PERCEPTION

Low Visual Similarity (Blur)

Original Observation o_t

Reconstruction $\hat{o}_t$ (MSE ≈ 50)

PROVOCATIVE ANALOGY: HUMAN DRIVING



o_t

HUMAN WORLD MODEL

1. FUNCTIONAL IMAGINATION (Rough prediction) ✓

predict function, dynamics, key objects

2. PIXEL-PERFECT RECONSTRUCTION (Must reconstruct $\hat{o}_t$) ✗

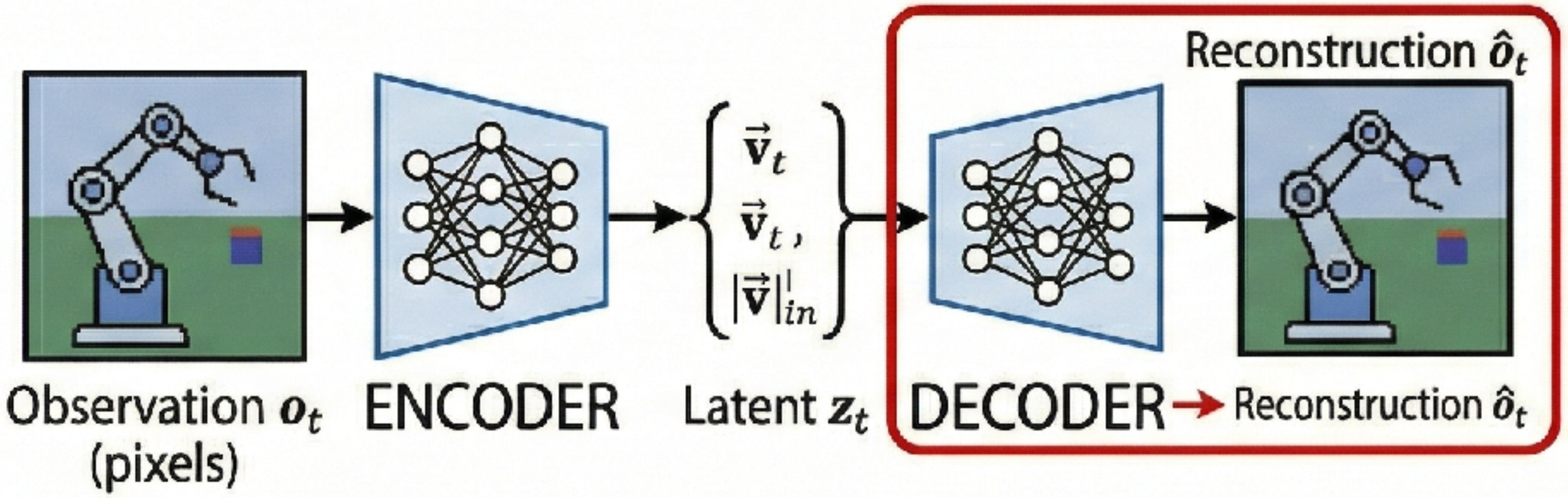
predict thousands of leaves to minimize MSE

INSIGHT: A GOOD WORLD MODEL PREDICTS THE *FUNCTION*, NOT THE *APPEARANCE*.

IGNORE IRRELEVANT DETAILS (e.g., LEAVES, TEXTURES). FOCUS ON TASK-RELEVANT INFORMATION.

SUMMARY: THE PARADIGM SHIFT

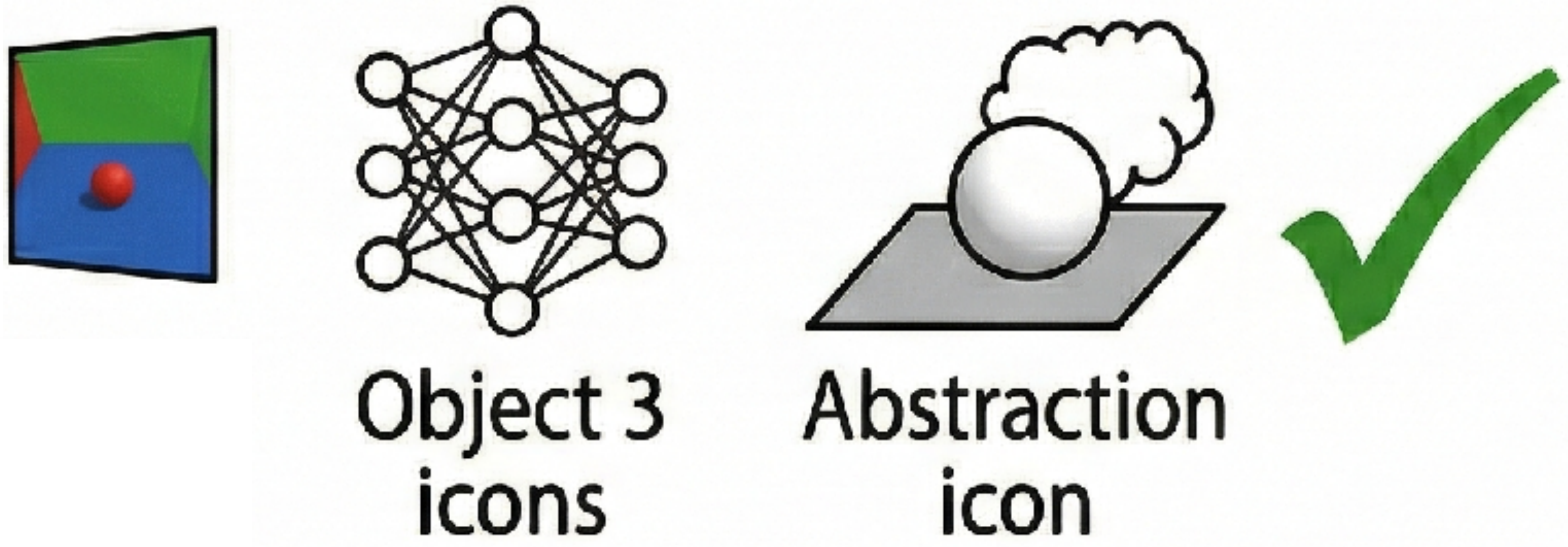
PIXEL-SPACE WORLD MODEL


Reconstruction $\hat{o}_t$
$\vec{v}_t$
$\vec{v}_t$,
$|\vec{v}|_{in}$
Observation o_t (pixels)
ENCODER
Latent z_t
DECODER → Reconstruction $\hat{o}_t$

✗

- High Reconstruct Cost
- Metric = MSE/Pixels
- Focus on *Appearance*
- Data Hungry

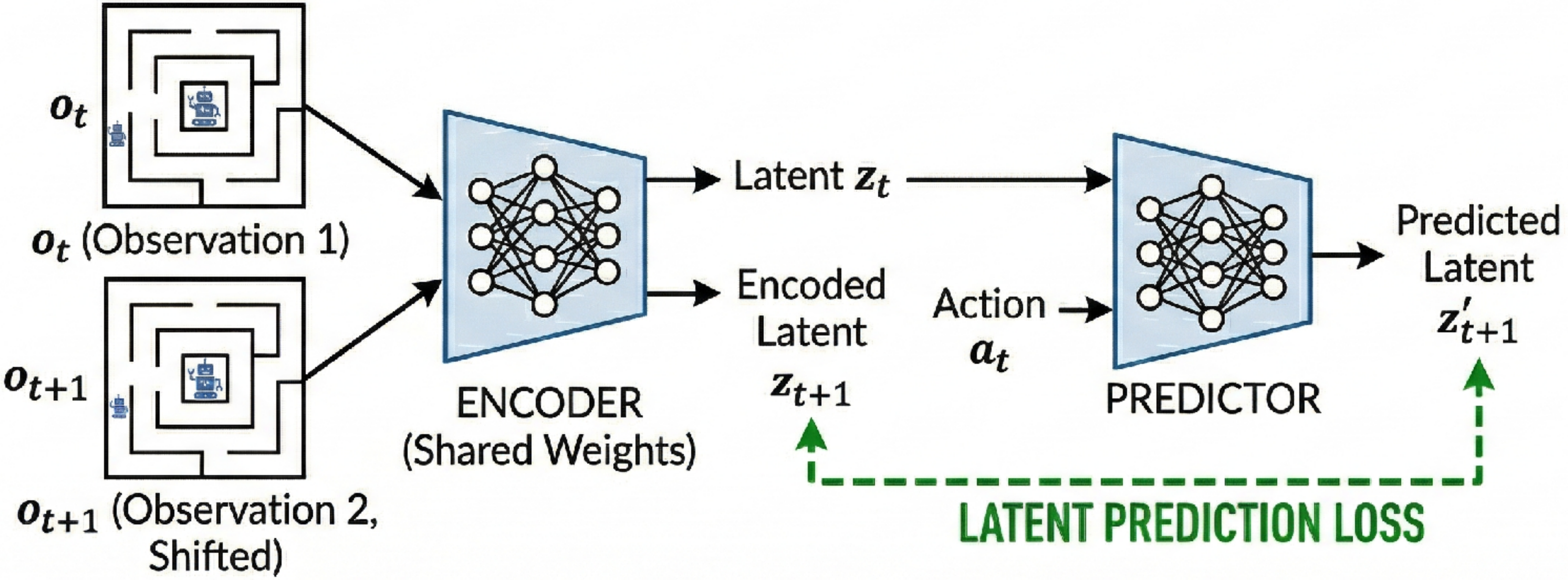
RECONSTRUCTION-FREE WORLD MODEL



✓

- ✓ Zero Reconstruct Cost
- Metric = Dynamics/Function
- Focus on *Structure*
- Data Efficient

JOINT-EMBEDDING PREDICTIVE ARCHITECTURE (JEPA)



WILL THAT WORK OUT OF THE BOX?

Consider: Ambiguous Futures? Degenerate Solutions? Metric in Latent Space?

Thank you!
See you Friday!