ICLR Oral

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY
EIAS 东方理工高等研究院 EASTERN INSTITUTE FOR ADVANCED STUDY
華東師範大學 EAST CHINA NORMAL UNIVERSITY

Open-World Reinforcement Learning over Long Short-Term Imagination

Jiajian Li*, Qi Wang*, Yunbo Wang†,

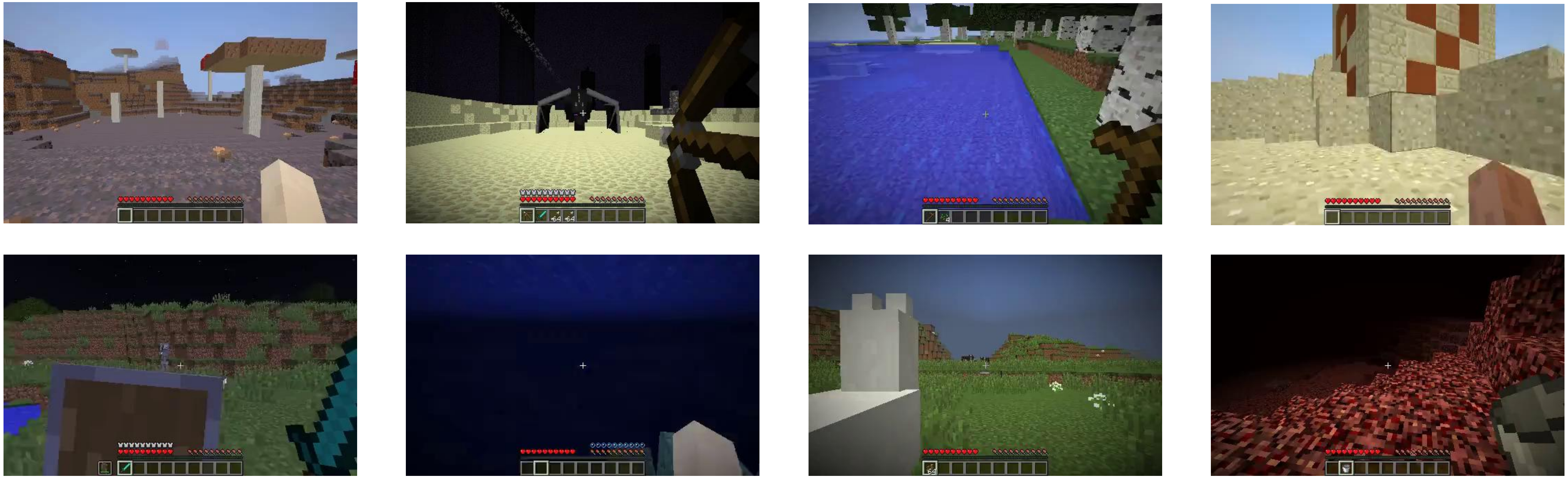
Xin Jin, Yang Li, Wenjun Zeng, Xiaokang Yang

*Equal contribution †Corresponding author

Motivation

Open-World RL Challenges

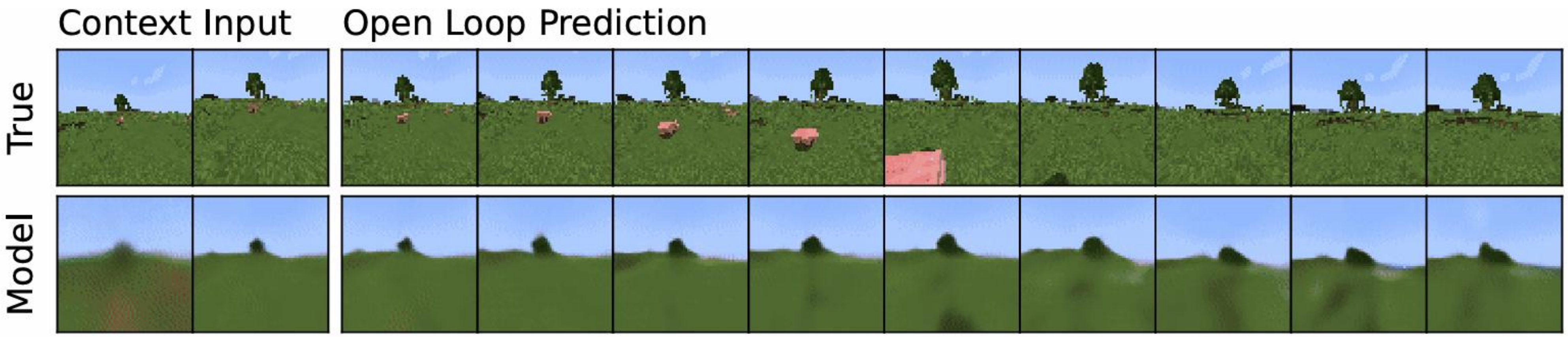
- Agents operate in large, dynamic environments with vast state spaces
- Policies must be highly flexible to interact with various objects and tasks
- Agents perceive the world with uncertainty, relying on raw visual input



Motivation

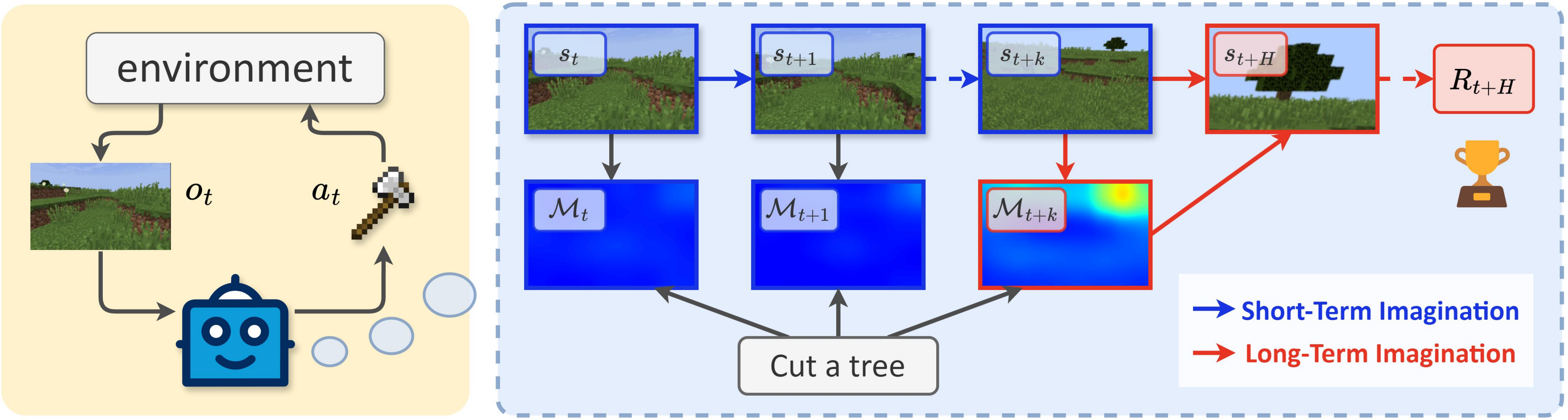
Limitations of Existing Methods

- Existing methods like *Voyager*[1] rely on handcrafted APIs, limiting real-world applicability
- Model-free RL methods like *DECKARD*[2] struggle with understanding environment mechanics and suffer from inefficient trial-and-error exploration
- Model-based RL methods like *DreamerV3*[3] improve sample efficiency but remain short-sighted, failing to explore vast solution spaces effectively



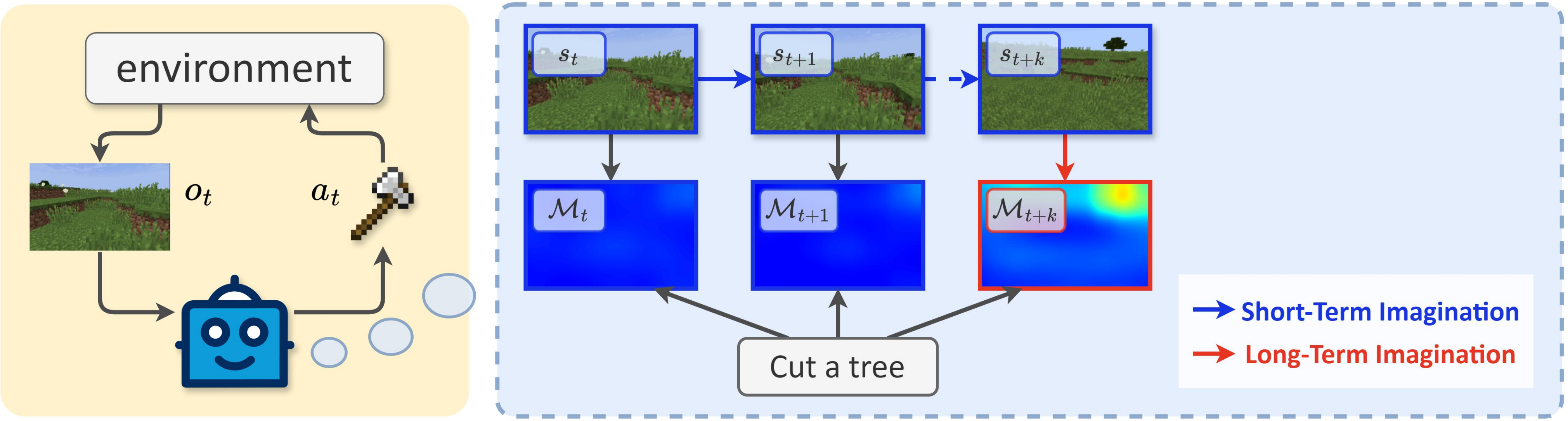
[1] Wang et al. *"Voyager: An Open-Ended Embodied Agent with Large Language Models."* TMLR, 2024.
[2] Nottingham et al. *"Do Embodied Agents Dream of Pixelated Sheep: Embodied Decision Making Using Language Guided World Modelling."* ICML, 2023.
[3] Hafner et al. *"Mastering Diverse Domains through World Models."* arXiv preprint arXiv:2301.04104, 2023.

Long Short-Term Imagination (LS-Imagine)

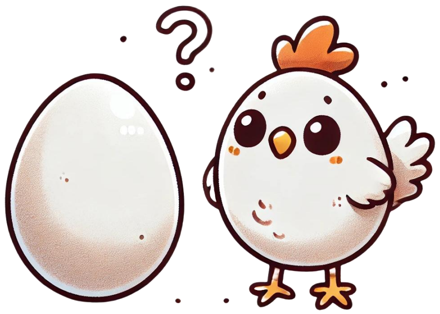


- Enable the world model to efficiently simulate the long-term effects of specific behaviors without the need for repeatedly rolling out one-step predictions
- Once trained, the long short-term world model provides both instant and jumpy state transitions along with corresponding (intrinsic) rewards, facilitating policy optimization in a joint space of short- and long-term imaginations

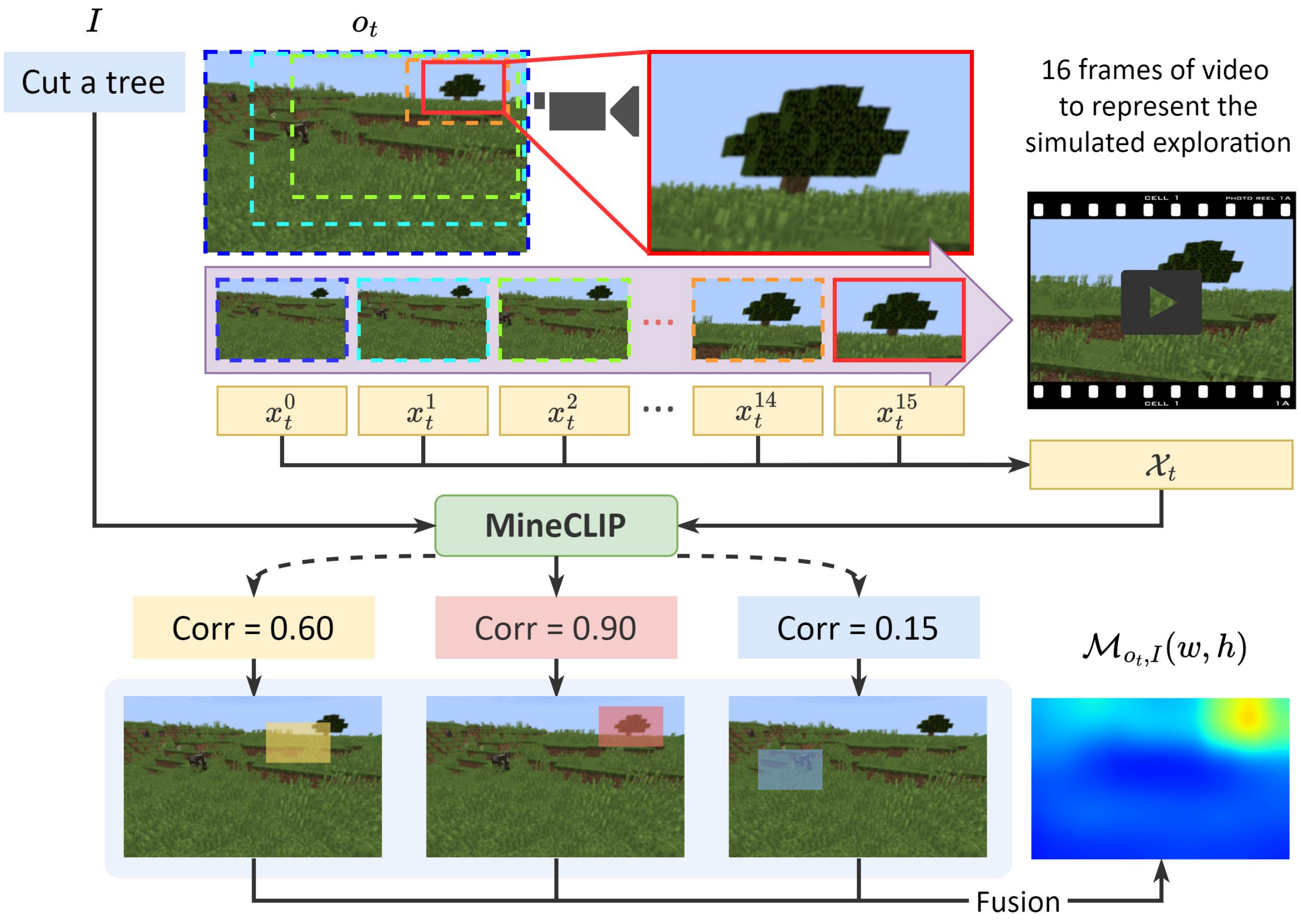
"Chicken-and-Egg" Dilemma



- Without true data showing the agent has reached the goal, how can we effectively train the model to simulate jumpy transitions from current states to pivotal future states that suggest a high likelihood of achieving that goal?



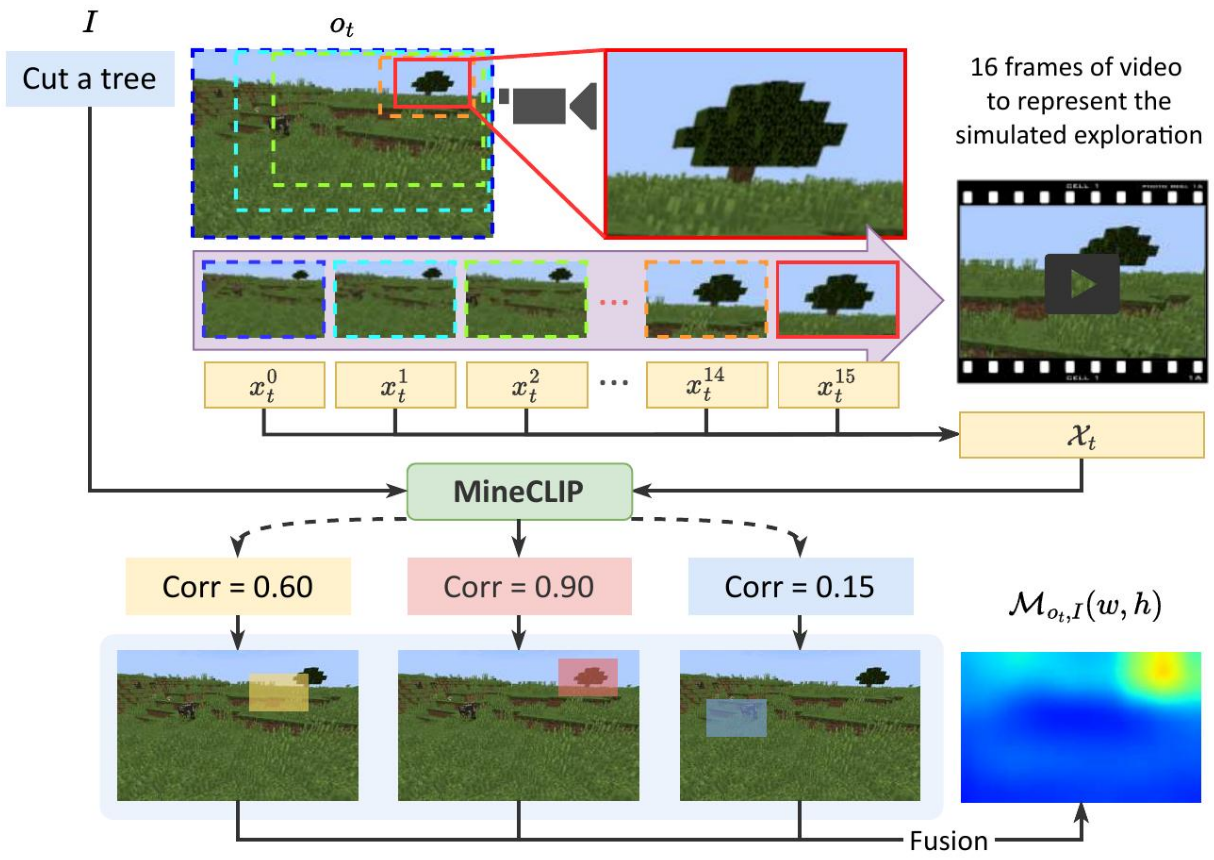
Affordance Map[4] Generation

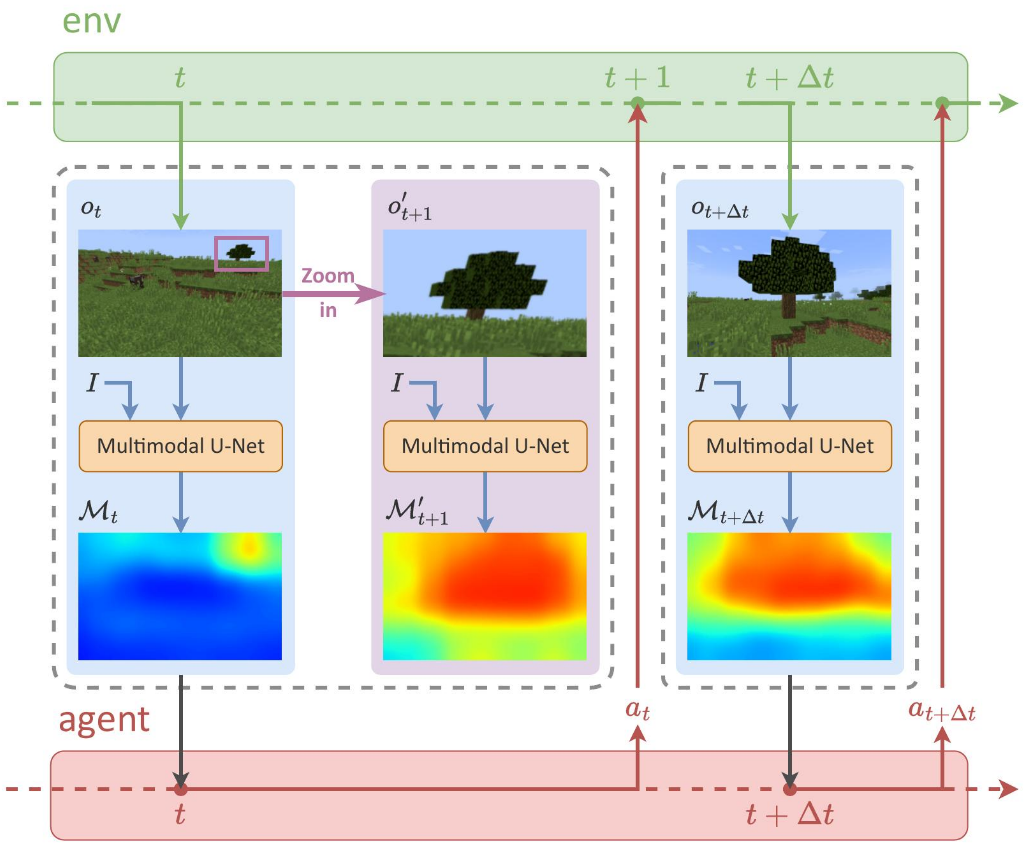


- Employ a sliding bounding box to scan individual images
- Execute continuous zoom-ins inside the bounding box
- Assess the relevance of the fake video clips to task-specific goals expressed in text using MineCLIP[5] model
- Fuse the relevance values at each bounding box position to generate a comprehensive affordance map

[4] Qi et al. *"Learning to Move with Affordance Maps."* ICLR, 2020.
[5] Fan et al. *"MineDojo: Building Open-Ended Embodied Agents with Internet-Scale Knowledge."* NeurIPS, 2022.

Rapid Affordance Map Generation

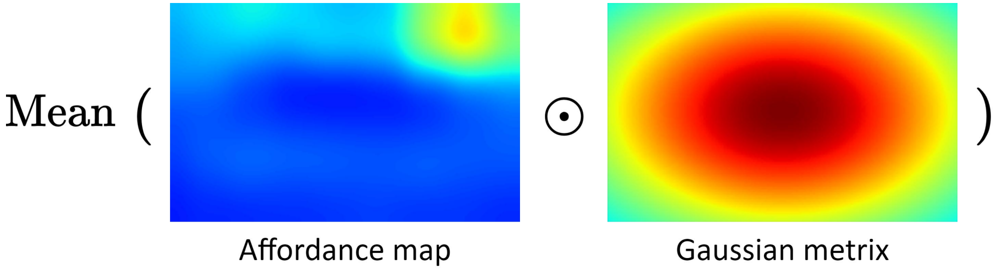


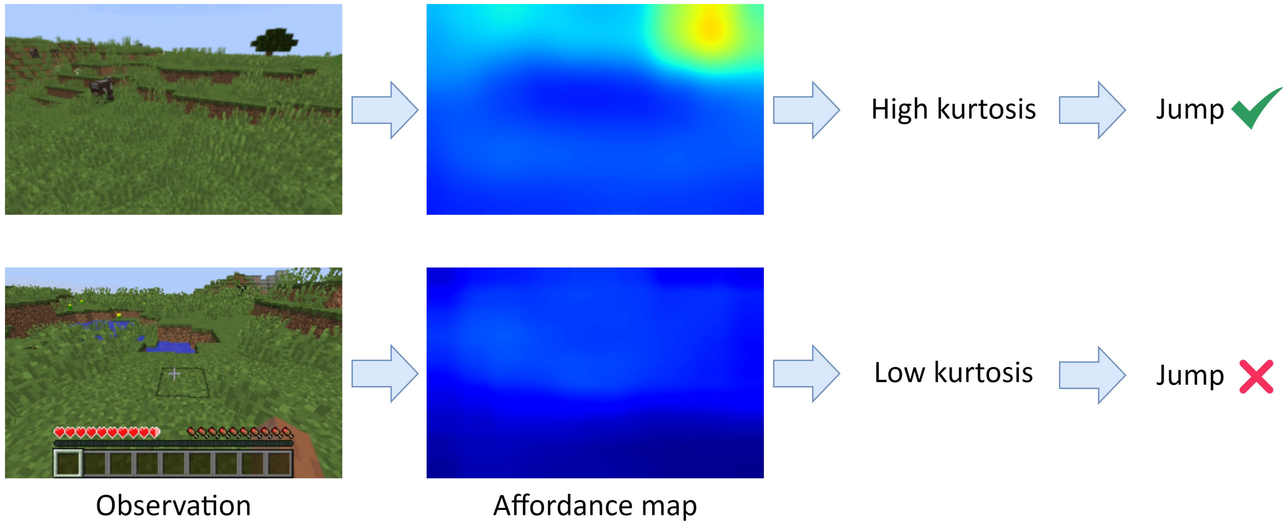


- Train a multimodal U-Net module[6] to approximate the affordance maps annotated through the proposed affordance map generation process for the sake of efficiency

[6] Cao et al. *"Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation."* ECCVW, 2022.

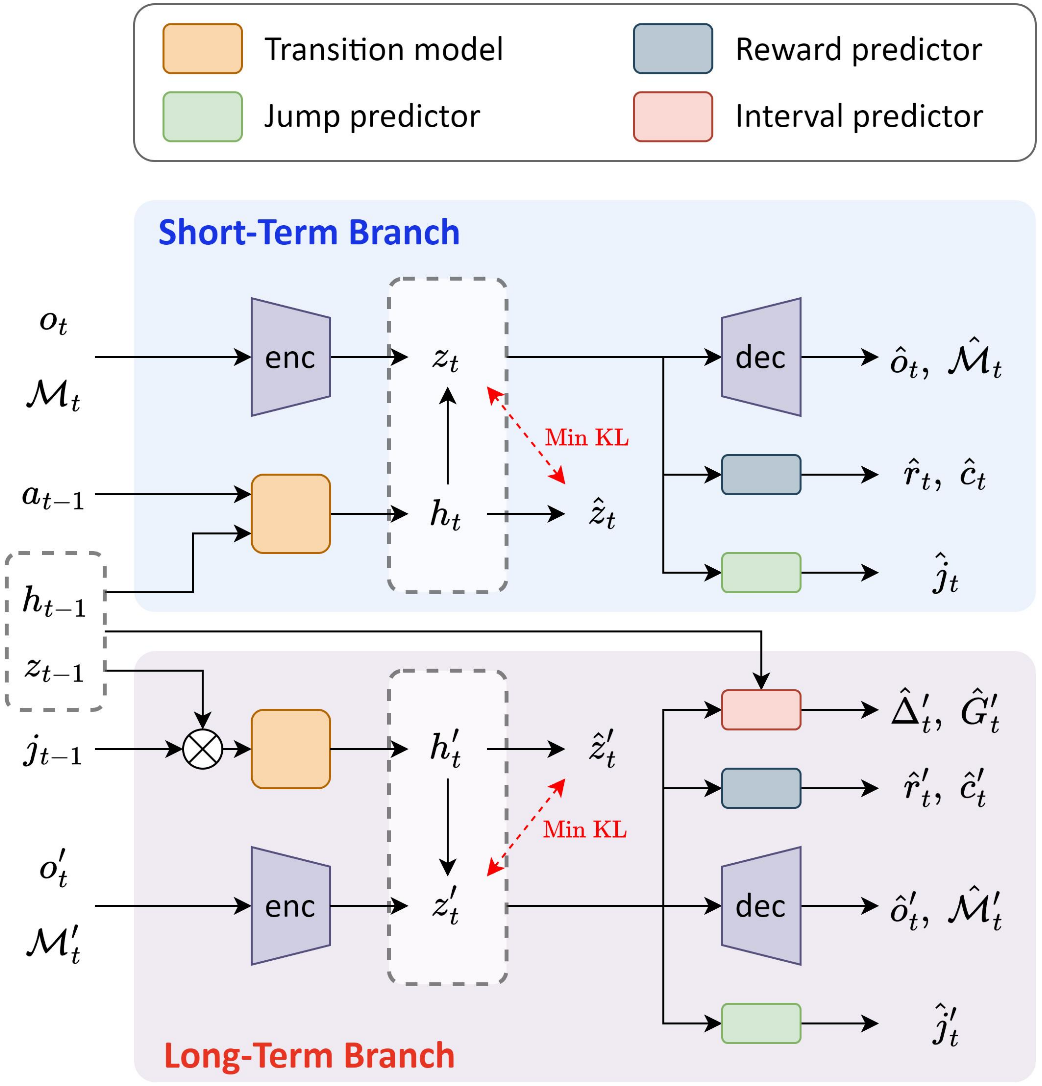
Affordance-Based Intrinsic Reward and Jumping Flag





- Compute the mean of the element-wise product of the affordance map and a same-shaped 2D Gaussian matrix as the affordance-driven intrinsic reward
- When a distant task-related target appears in the agent's observation, which can be reflected by a higher kurtosis in the affordance map, a jumpy state transition should be adopted

Long Short-Term World Model



Short-term transition model: $h_t = f_\phi(h_{t-1}, z_{t-1}, a_{t-1})$

Long-term transition model: $h'_t = f_\phi(h_{t-1}, z_{t-1})$

Encoder: $z_t \sim q_\phi(z_t \mid h_t, o_t, \mathcal{M}_t)$

Dynamics predictor: $\hat{z}_t \sim p_\phi(\hat{z}_t \mid h_t)$

Reward predictor: $\hat{r}_t, \hat{c}_t \sim p_\phi(\hat{r}_t, \hat{c}_t \mid h_t, z_t)$

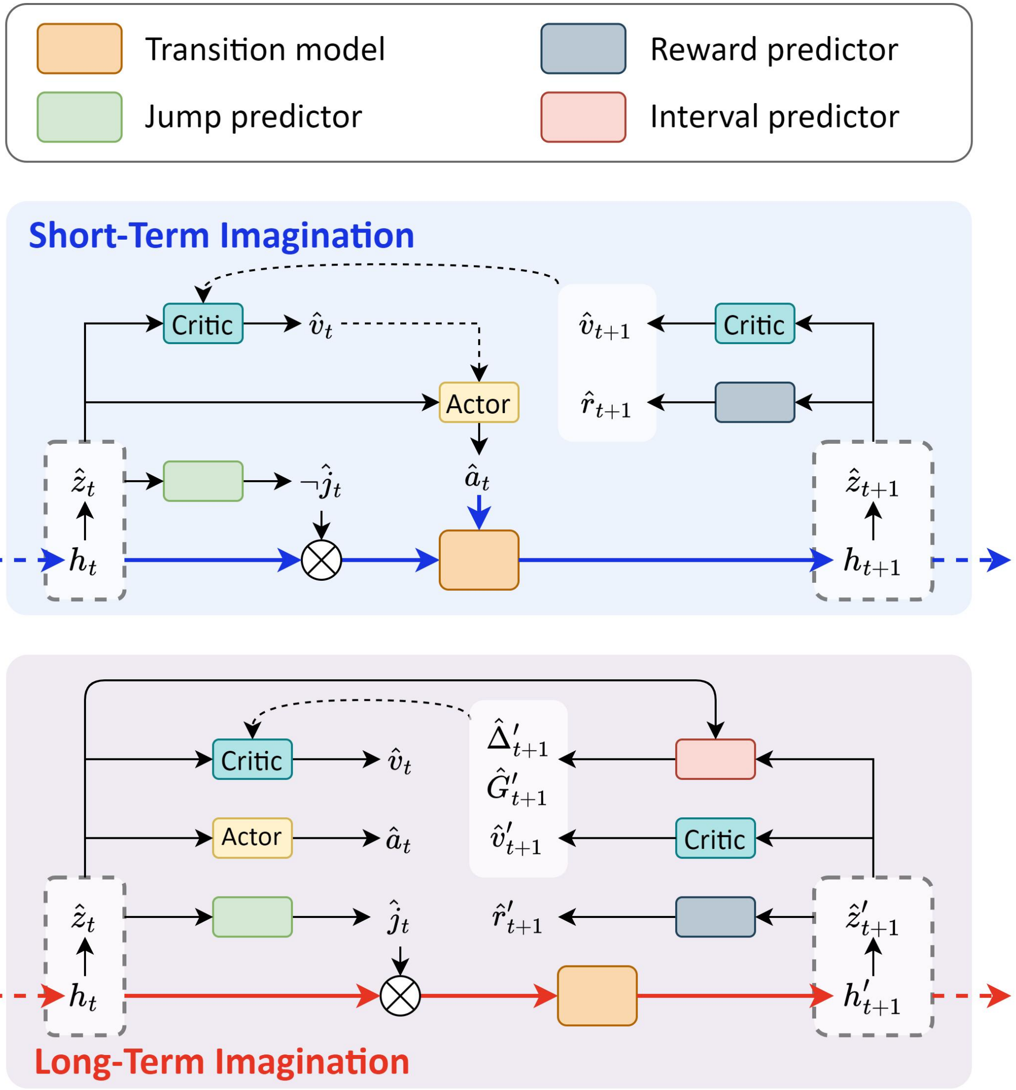
Decoder: $\hat{o}_t, \hat{\mathcal{M}}_t \sim p_\phi(\hat{o}_t, \hat{\mathcal{M}}_t \mid h_t, z_t)$

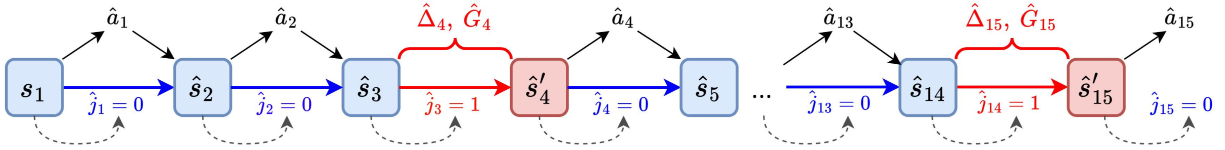
Jump predictor: $\hat{j}_t \sim p_\phi(\hat{j}_t \mid h_t, z_t)$

Interval predictor: $\hat{\Delta}_t{}', \hat{G}_t{}' \sim p_\phi(\hat{\Delta}_t{}', \hat{G}_t{}' \mid h_{t-1}, z_{t-1}, h'_t, z'_t)$

- The state transition model includes both short-term and long-term branches
- Use the affordance map as an input of the encoder, which serves as the goal-conditioned prior guidance to the agent

Behavior Learning over Mixed Long Short-Term Imagination





- Dynamically select either the long-term transition model or the short-term transition model to predict subsequent states based on the jumping flag predicted by the jump predictor

$$R_t^\lambda \doteq \begin{cases} \hat{c}_t\{\hat{G}_{t+1} + \gamma^{\hat{\Delta}_{t+1}}\left[(1-\lambda)v_\psi(\hat{s}_{t+1}) + \lambda R_{t+1}^\lambda\right]\} & \text{if } t < L \\ v_\psi(\hat{s}_L) & \text{if } t = L \end{cases}$$

- Employ an actor-critic algorithm to learn behavior from the latent state sequences predicted by the world model

Experiments

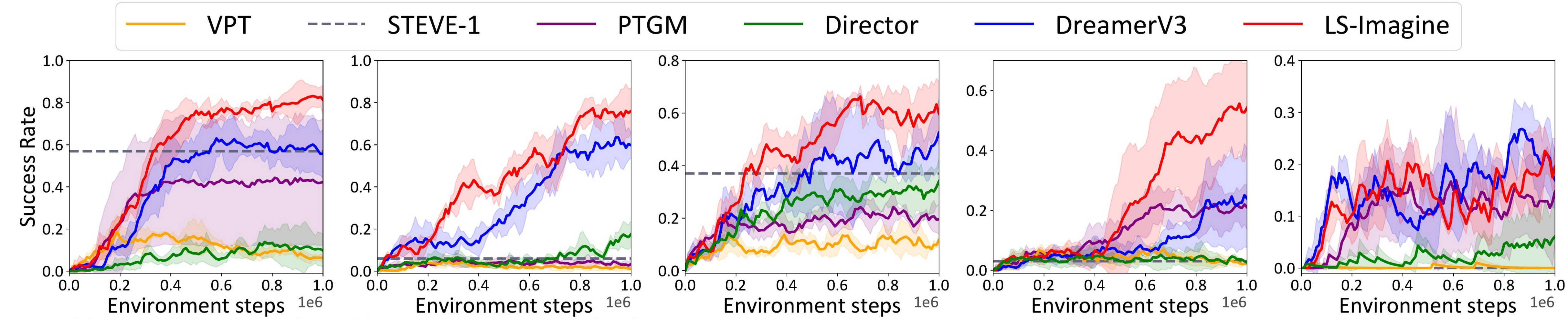
Table 1: Experimental setups of the Minecraft AI agents. *IL* is short for imitation learning.

Model	Controller	Observation	Video Demos
DECKARD (2023)	RL	Pixels & Inventory	✓
Auto MC-Reward (2024a)	IL + RL	Pixels & GPS	✗
Voyager (2024a)	GPT-4	Minecraft simulation & Error trace	✗
DEPS (2023)	IL	Pixels & Yaw/pitch angle & GPS & Voxel	✗
STEVE-1 (2023)	Generative model	Pixels	✗
VPT (2022)	IL + RL	Pixels	✓
DreamerV3 (2023)	RL	Pixels	✗
LS-Imagine	RL	Pixels	✗

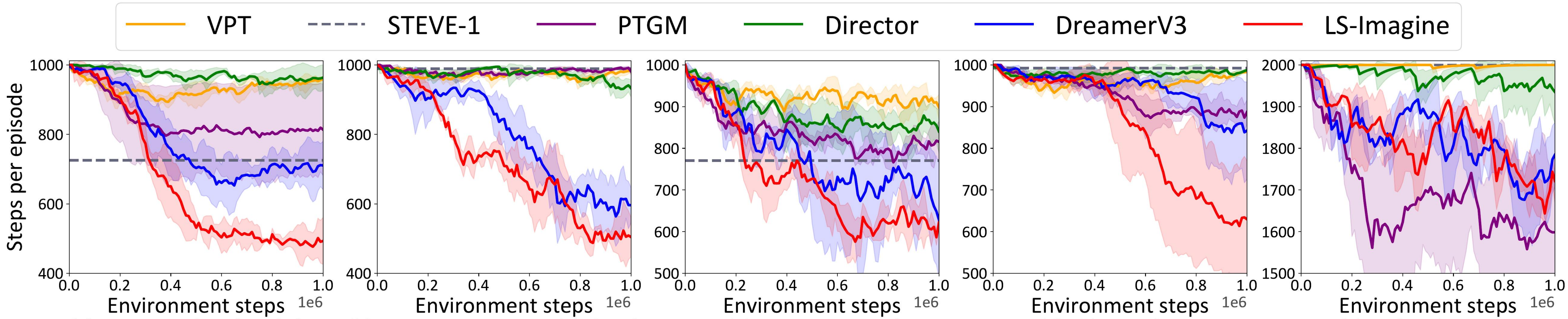
Table 2: Details of the MineDojo tasks.

Task	Language description	Initial tools	Initial mobs and distance	Max steps
Harvest log in plains	"Cut a tree."	–	–	1000
Harvest water with bucket	"Obtain water."	bucket	–	1000
Harvest sand	"Obtain sand."	–	–	1000
Shear sheep	"Obtain wool."	shear	sheep, 15	1000
Mine iron ore	"Mine iron ore."	stone pickaxe	–	2000

Results

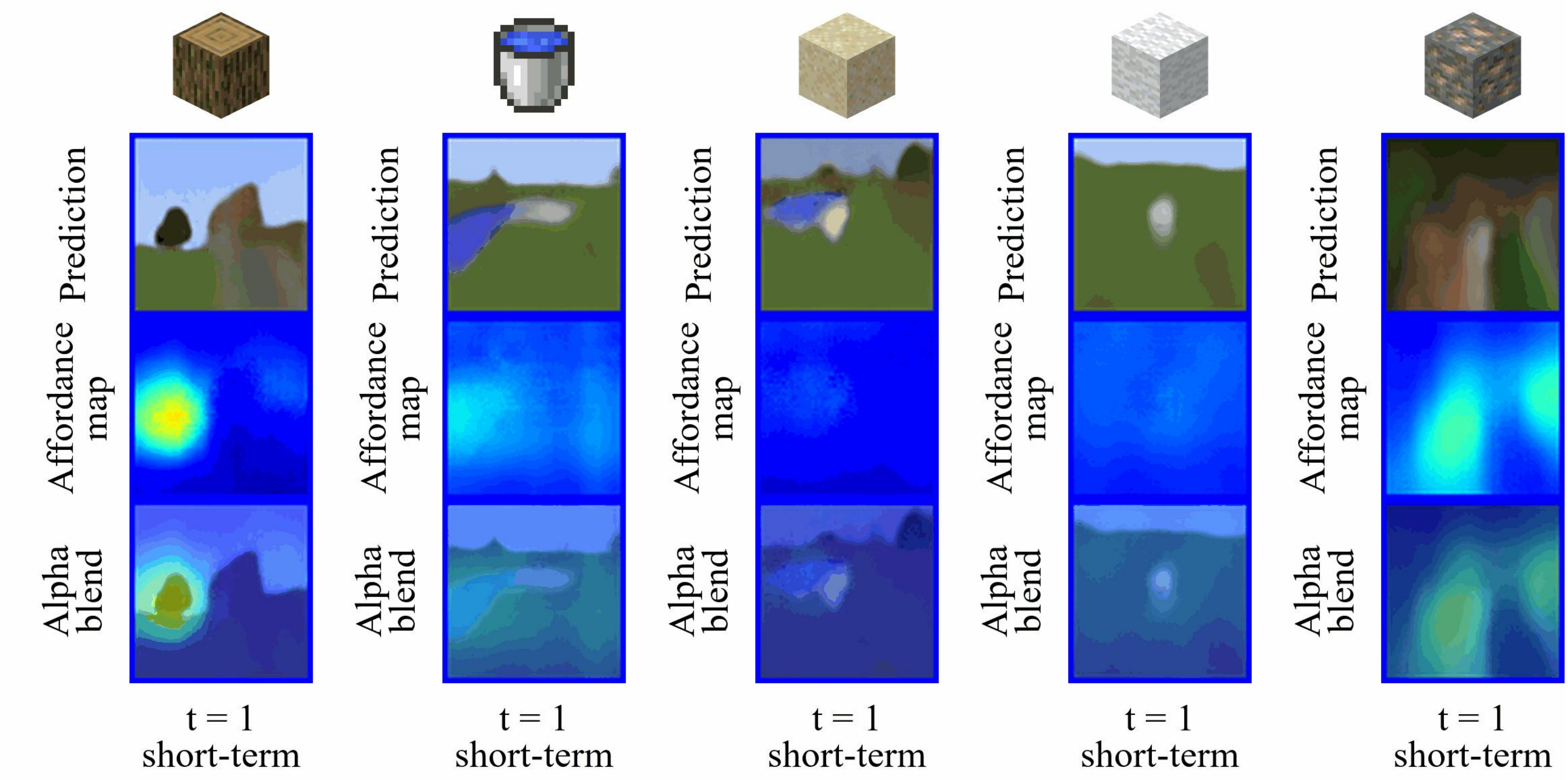


(a) Harvest log in plains (b) Harvest water with bucket (c) Harvest sand (d) Shear sheep (e) Mine iron ore



(a) Harvest log in plains (b) Harvest water with bucket (c) Harvest sand (d) Shear sheep (e) Mine iron ore

Visualization of the Long Short-Term Imaginations



Conclusion

- Extend the imagination horizon and leverage a long short-term world model to facilitate efficient off-policy exploration across expansive state spaces
- Incorporate goal-conditioned jumpy state transitions and affordance maps to help agents better grasp long-term value
- Enhance agents' decision-making abilities by improving their understanding of long-term value through structured exploration mechanisms

Oral:

- Oral Session 2A
- Thu 24 Apr 4:30 p.m. CST — 4:42 p.m. CST

Poster:

- Poster Session 1
- Poster Sessions Hall TBD, Thu 24 Apr 10 a.m. CST — 12:30 p.m. CST

https://qiwang067.github.io/ls-imagine