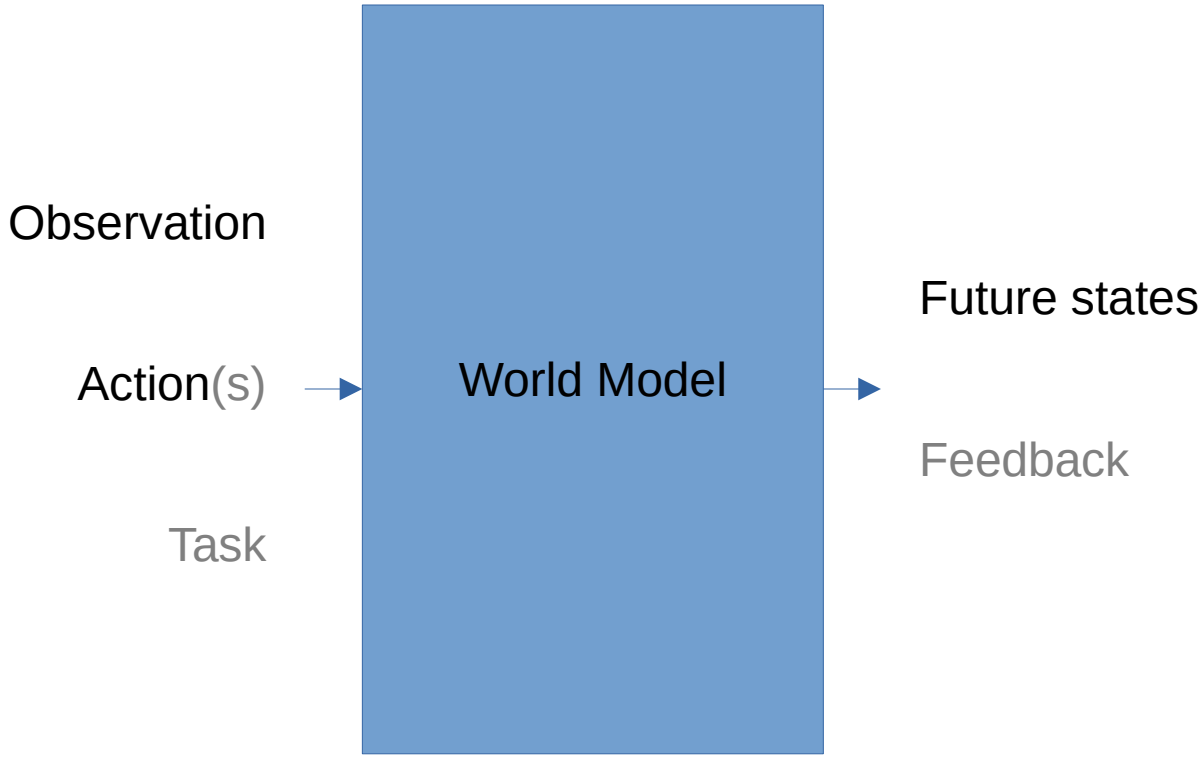
World Model as Simulator

Thomas Vitry

http://www.informatik.uni-hamburg.de/WTM/

In Essence



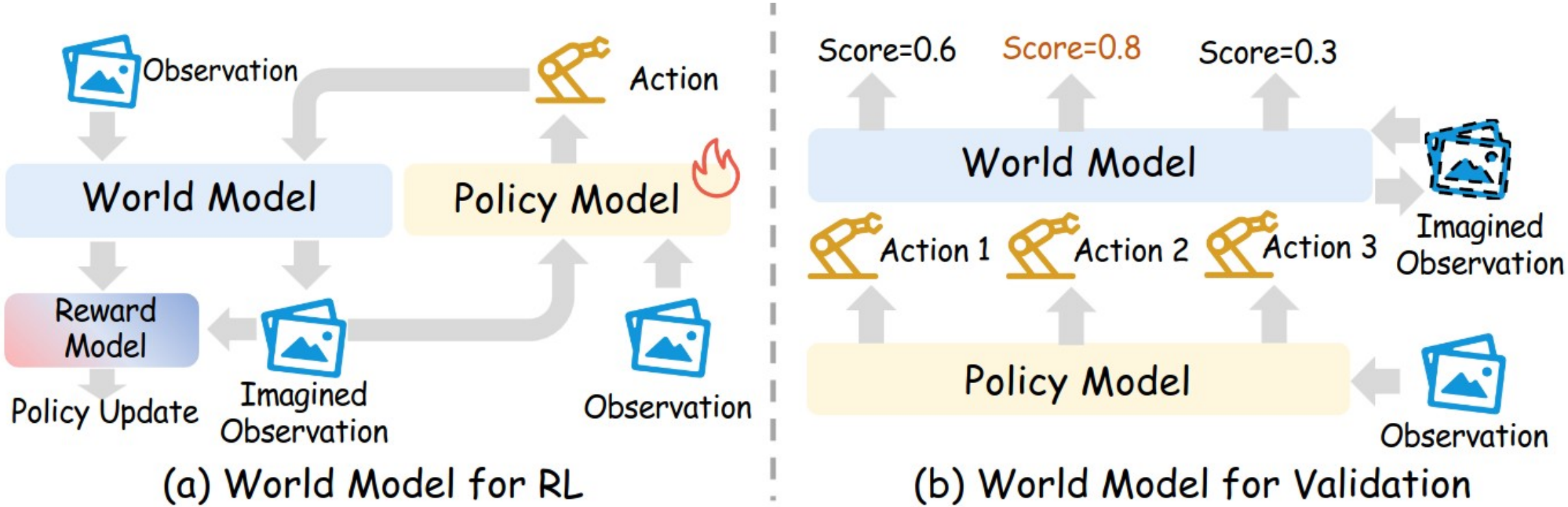
KNOWLEDGE TECHNOLOGY

Motivation

- Reinforcement Learning in the real world:
  - Slow, expensive
  - Difficult to reset
  - Potentially unsafe
- Imitation Learning:
  - Low demonstration quality
  - Cannot learn from failure
- RL in Simulation:
  - Fast, safe and easy to reset
  - Still somewhat expensive
  - Does not port well to real
- RL in World Model:
  - Like RL in Simulation
  - Can be conditioned to real world
  - Cheap

KNOWLEDGE TECHNOLOGY

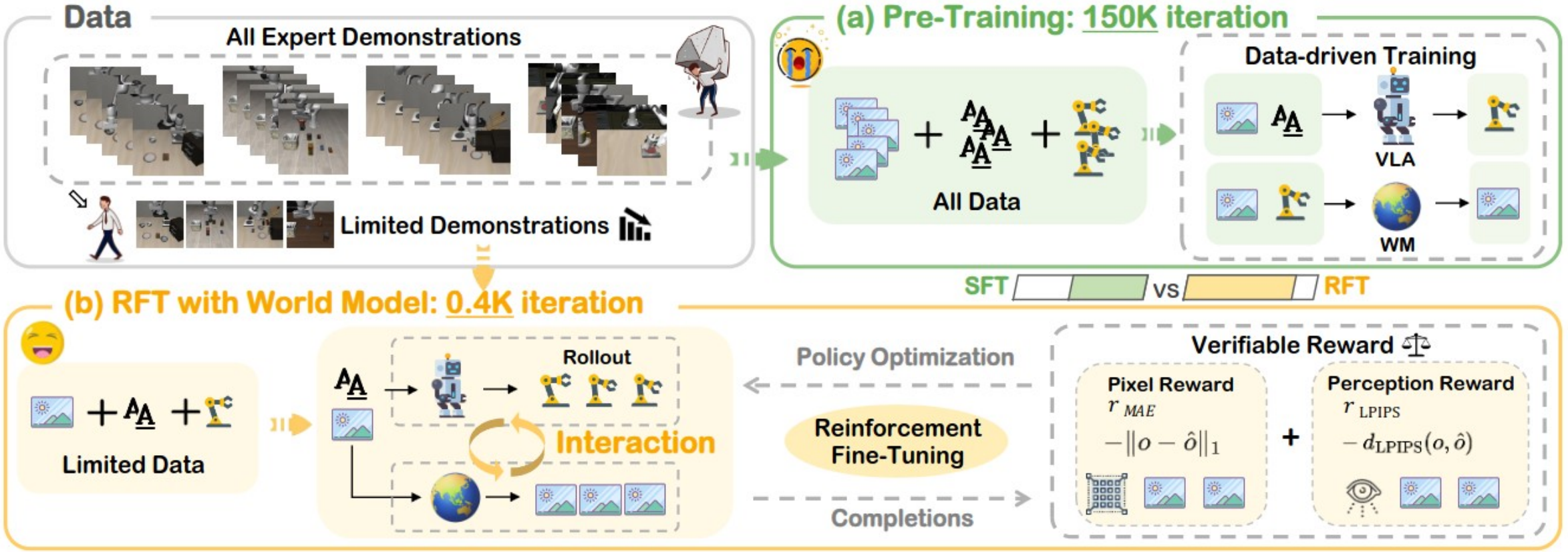
Two use cases



[1] Survey

KNOWLEDGE TECHNOLOGY

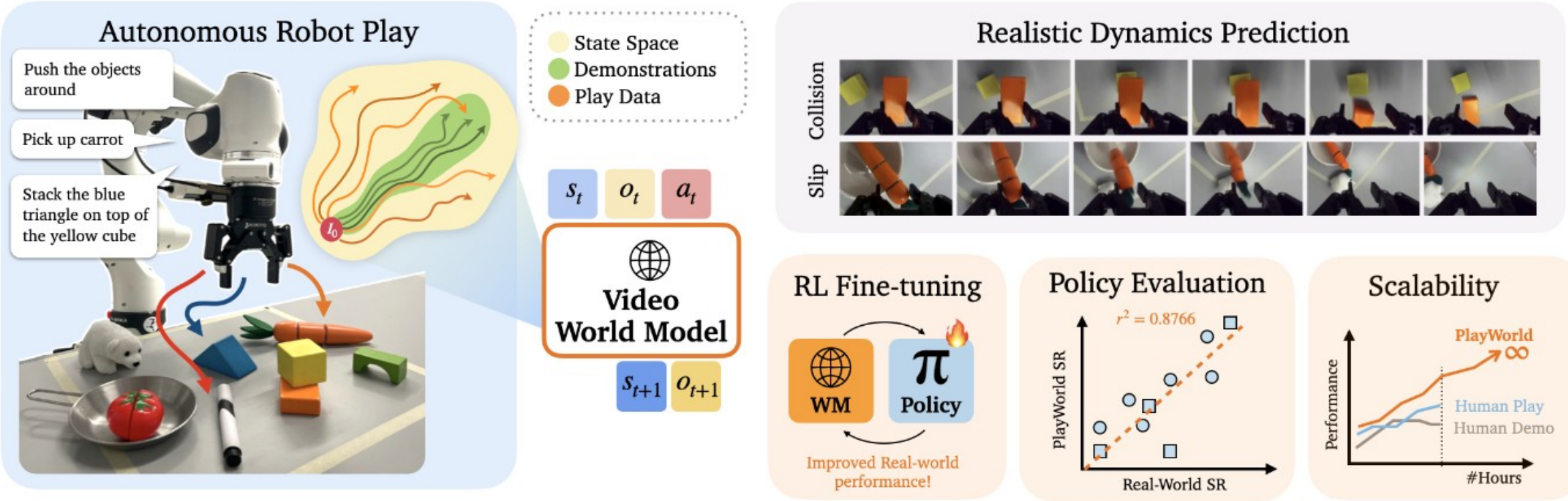
WM for RL
V1 – Train the policy in imagination



[2] VLA-RFT

KNOWLEDGE TECHNOLOGY

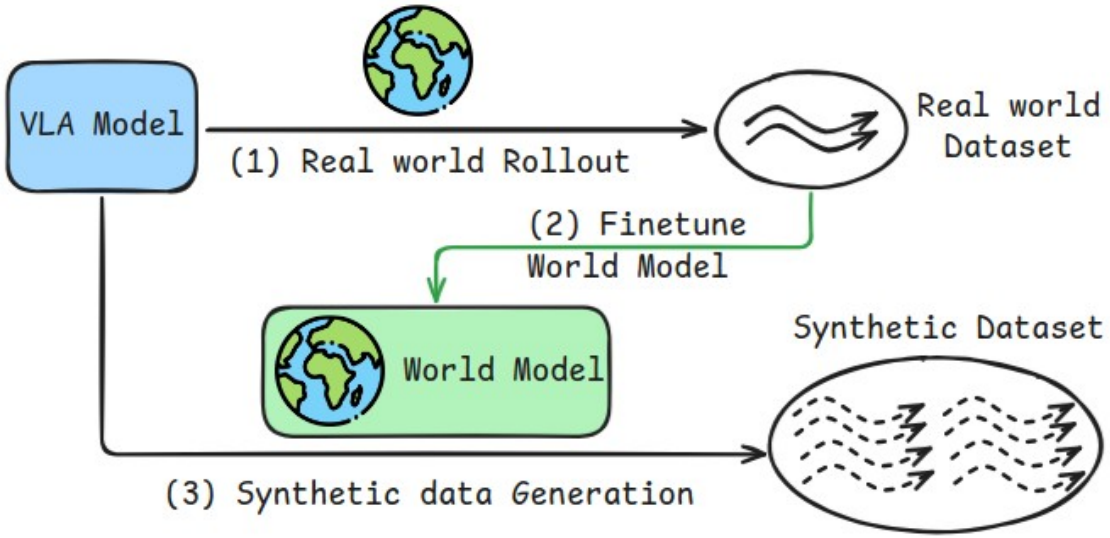
WM for RL
V2 – Train the WM through exploration



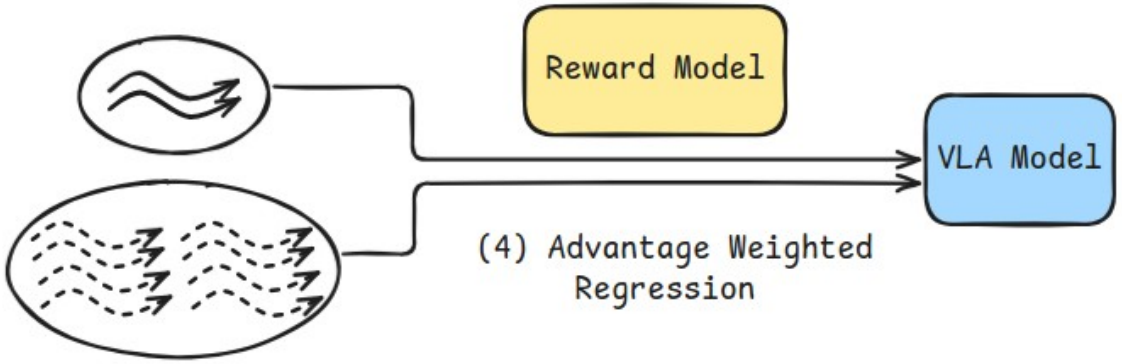
[3] PlayWorld

KNOWLEDGE TECHNOLOGY

WM for RL
V3 – Train the WM and policy together



1. Learning accurate world model

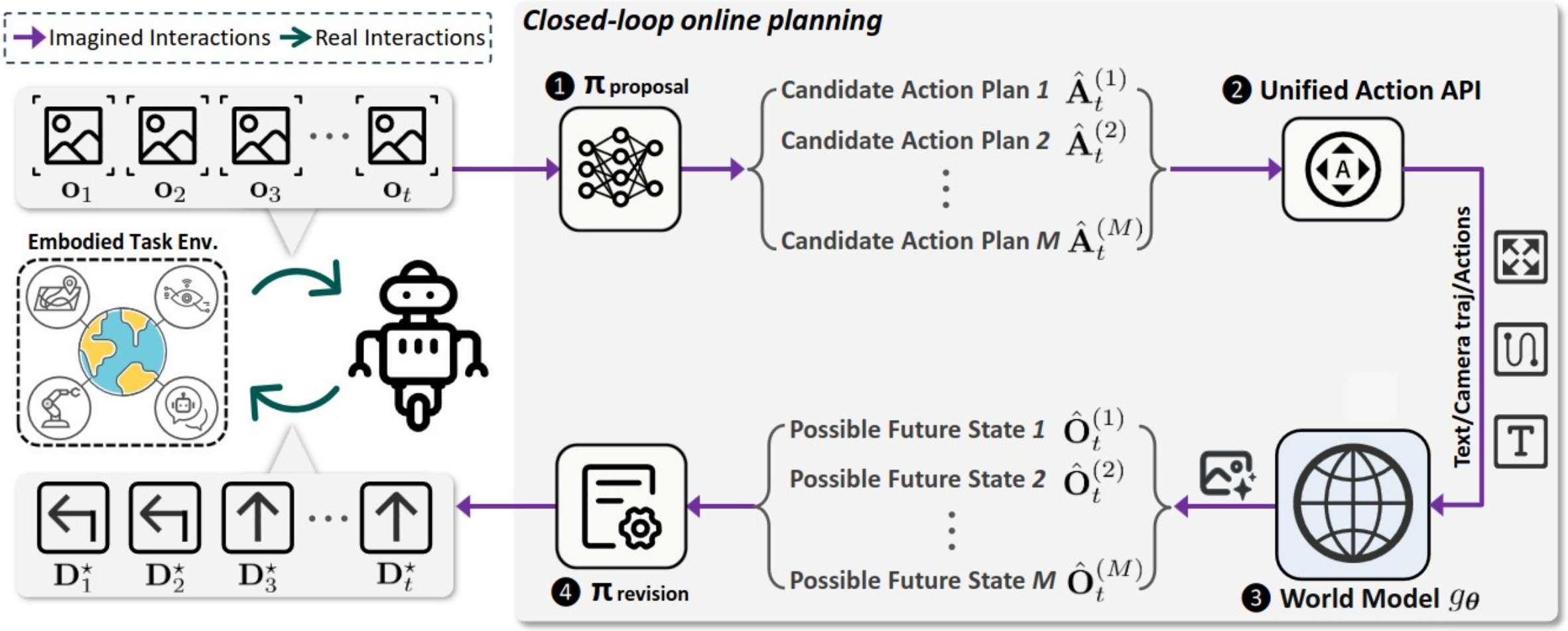


2. Regularized RL with Expressive Policy

[4] VLAW

KNOWLEDGE TECHNOLOGY

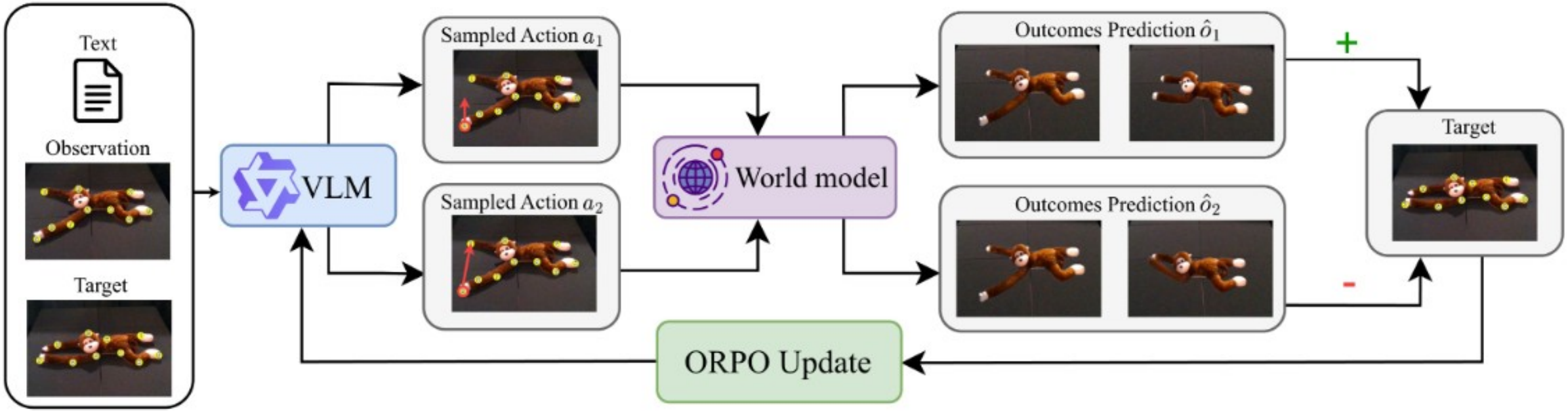
WM for Validation
V1 – The WM as plan evaluator



[5] World-in-World

KNOWLEDGE TECHNOLOGY

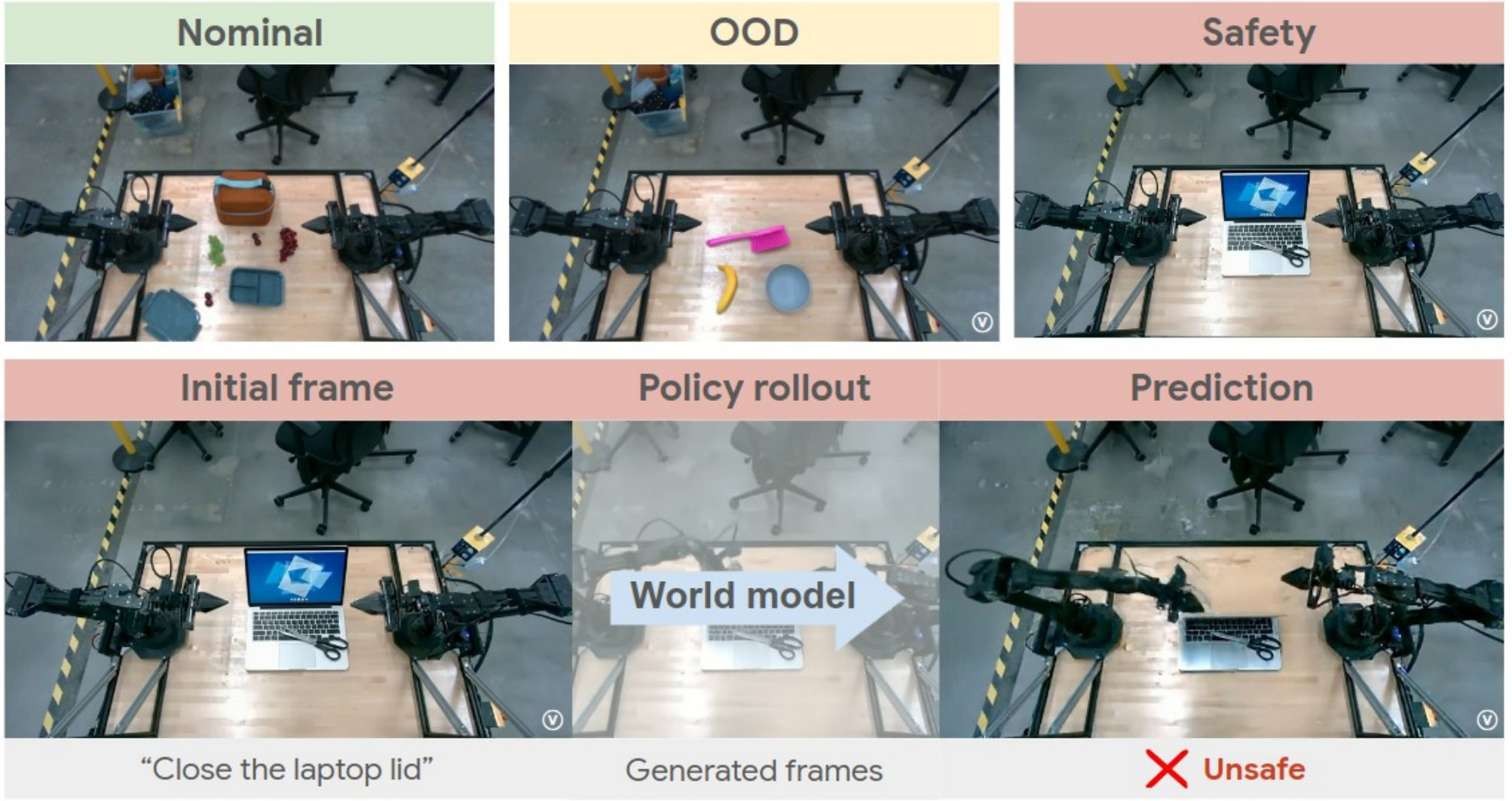
WM for Validation
V2 – Mix of evaluation and RL



[6] DreamPlan

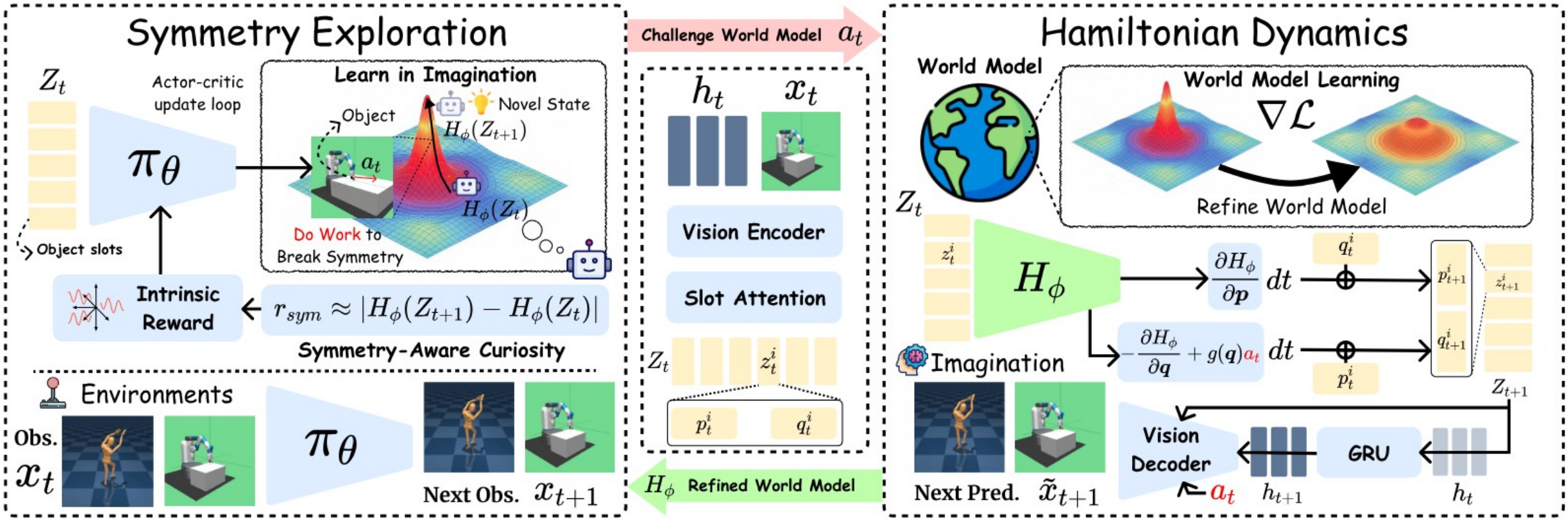
KNOWLEDGE TECHNOLOGY

WM for Validation
WM as pure evaluation environment



[7]

DreamSAC: An interesting approach



KNOWLEDGE TECHNOLOGY

DreamSAC: WM Formulas

- Hamiltonian Dynamics

$$\frac{d\boldsymbol{q}}{dt} = \frac{\partial H_\phi(z)}{\partial \boldsymbol{p}}, \quad \frac{d\boldsymbol{p}}{dt} = -\frac{\partial H_\phi(z)}{\partial \boldsymbol{q}} + g(\boldsymbol{q})a_t$$

- Invariant to transform

$$H_\phi(g \cdot Z_t) = H_\phi(Z_t), \quad \forall g \in G$$

- Loss funtion (ELBO)

$$\mathcal{L}_{\text{total}}(\phi) = \sum_{t=1}^{T} \Big(\mathcal{L}_{\text{pred}}(\phi) + \beta_{\text{dyn}} \mathcal{L}_{\text{dyn}}(\phi) + \beta_{\text{rep}} \mathcal{L}_{\text{rep}}(\phi) + \gamma \mathcal{L}_{\text{vr}}(\phi) \Big)$$

- Constrastive Viewpoint Robustness Loss

$$\mathcal{L}_{\text{vr}}(\phi) = -\mathbb{E}\left[\log \frac{\exp(\text{sim}(Z_t^A, Z_t^B)/\tau)}{\sum_{j=1}^{K} \exp(\text{sim}(Z_t^A, Z_j^B)/\tau)} \right]$$

KNOWLEDGE TECHNOLOGY

DreamSAC: WM Exploration

- Intrinsic reward $r_{\text{sym},t+1} = \underbrace{|H_\phi(Z_{t+1}) - H_\phi(Z_t)|}_{\text{Symmetry Probing}} - \underbrace{\lambda_s ||a_t - a_{t-1}||^2}_{\text{Action Smoothness}}$
- Additional stability $r_{\text{int},t+1} = (1 - w_t) \cdot r_{\text{RND},t+1} + w_t \cdot r_{\text{sym},t+1}$

The End

Thank you for your attention.

Any question?

Literature:

- [1] Hou, Bohan, et al. "World Model for Robot Learning: A Comprehensive Survey." (2026)
- [2] Li, Hengtao, et al. "Vla-rft: Vision-language-action reinforcement fine-tuning with verified rewards in world simulators." (2025)
- [3] Yin, Tenny, et al. "Playworld: Learning robot world models from autonomous play." (2026)
- [4] Guo, Yanjiang, et al. "Vlaw: Iterative co-improvement of vision-language-action policy and world model." (2026)
- [5] Zhang, Jiahan, et al. "World-in-world: World models in a closed-loop world." (2025)
- [6] Jia, Emily Yue-Ting, et al. "DreamPlan: Efficient Reinforcement Fine-Tuning of Vision-Language Planners via Video World Models." (2026)
- [7] Team, Gemini Robotics, et al. "Evaluating Gemini Robotics Policies in a Veo World Simulator." (2025)
- [8] DreamSAC: Learning Hamiltonian World Models via Symmetry Exploration, Jinzhou Tang, Fan Feng, Minghao Fu, Wenjun Lin, Jing Yang, Biwei Huang, Keze Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec 2026

KNOWLEDGE TECHNOLOGY