Intern1의 논문 리뷰 2 Nov. 2020.

Plan2Explore

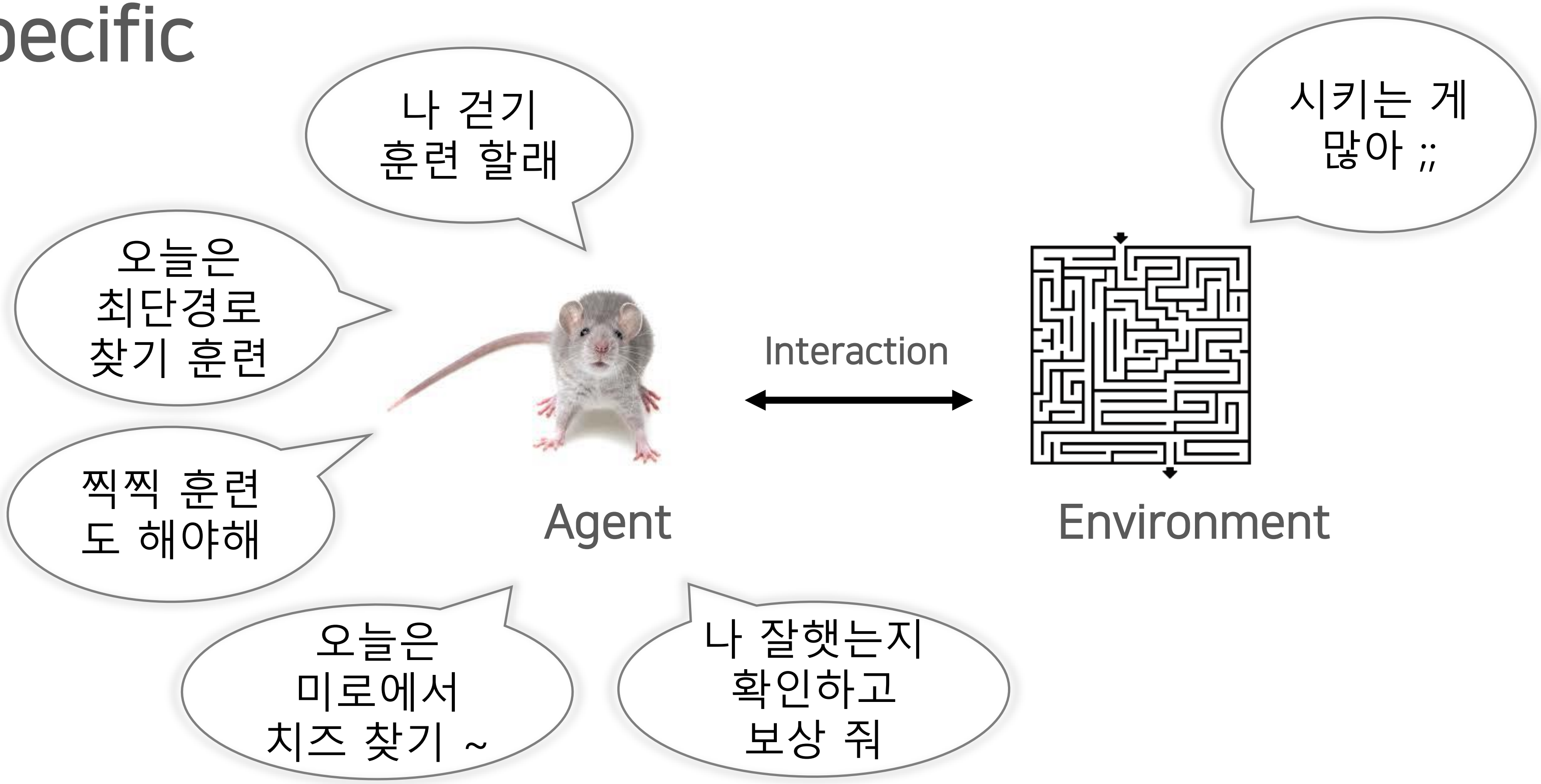
Planning to Explore via Self-Supervised World Models

Plan2Explore

- Self Supervised Model Based Learning
- Work directly from images
- Interact with environment without agent, to collect new data
- Train the World Model with Intrinsic Motivation

Supervised Learning

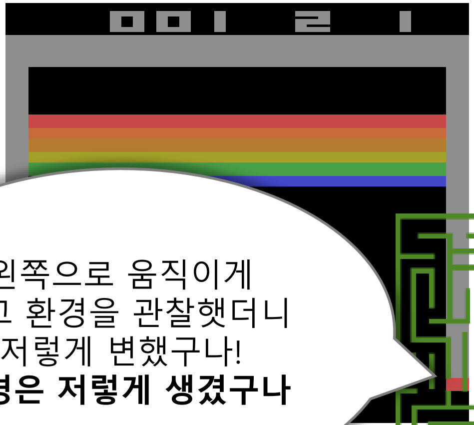
■ Task Specific



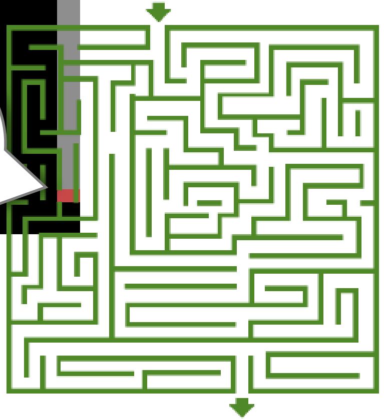
Number of Task ↑ → Require Large Amount of Experience

Self Supervised Model Based Learning

■ Train World Model



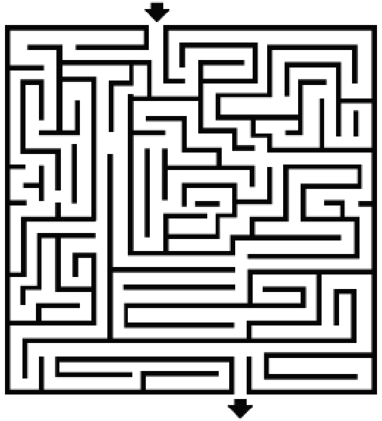
왼쪽으로 움직이게
하고 환경을 관찰했더니
저렇게 변했구나!
환경은 저렇게 생겼구나



Model

High Dimension
Interaction

Benefit for
High Dimension !



Environment

쥐돌이야
내가 대답해
줄게

Latent Space
Interaction

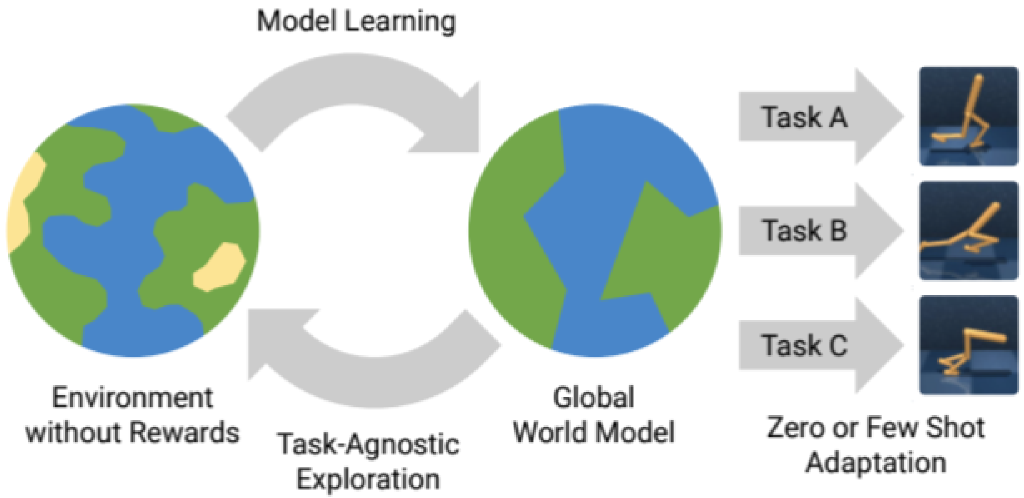
Fast

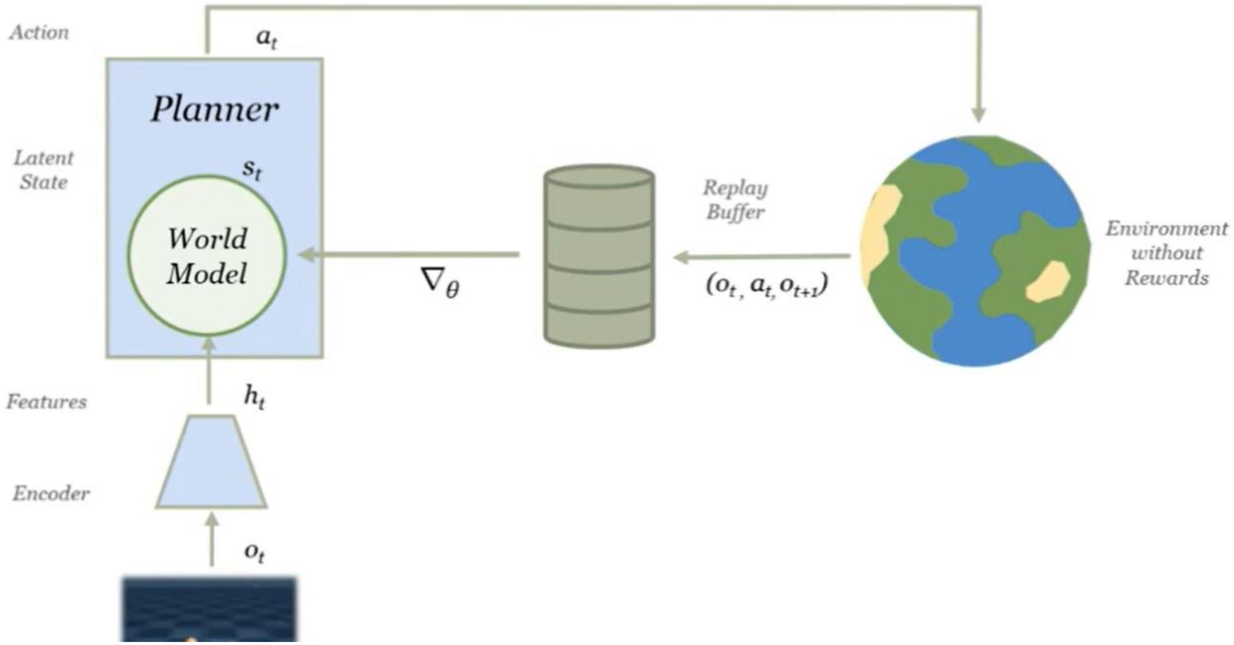
Model
Learn Task Independent Feature

Agent
Fast Adaptation to New Task

Exploration

Exploration





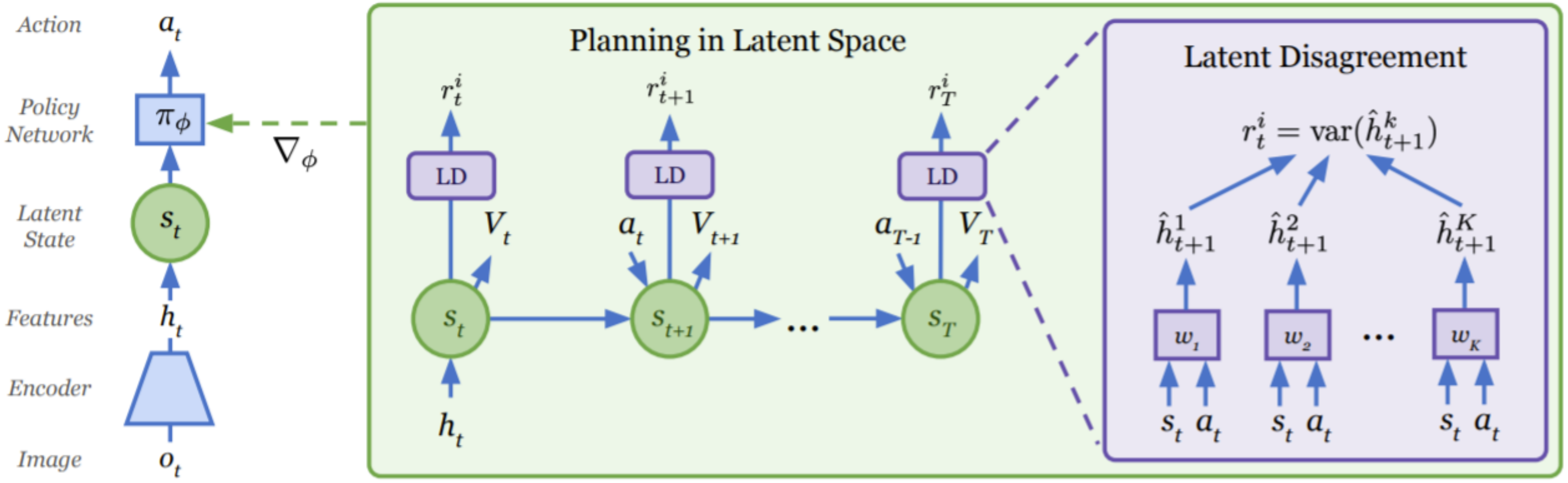
> Environment과의 Interaction으로 World Model 학습

> Planner로 다음 Interaction 결정

Key Idea Use long-term planning to collecting novel data

Exploration

■ Planning in Latent Space



Latent State를 이용하는
가벼운 1-step model들을 Ensemble

Motivation

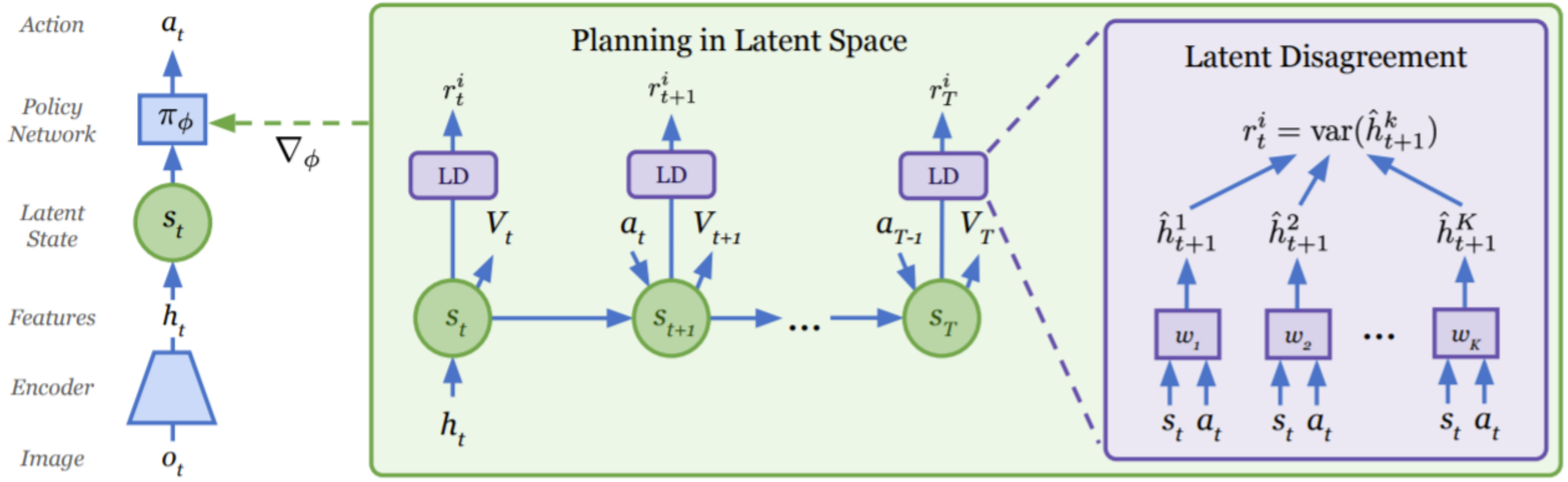
직관적으로, 초기화도 다르고, 관찰 순서도 다른 앙상블은 처음에는 예측이 다르나,
데이터가 늘어날수록 모델이 동일한 예측으로 수렴하며 불일치가 감소한다.

Exploration

■ Stochastic Model -> Ensemble

Exploration

■ Planning in Latent Space



Latent State를 이용하는
가벼운 1-step model들을 Ensemble

Ensemble predictors: $q(h_{t+1} \mid w_k, s_t, a_t)$

$$q(h_{t+1} \mid w_k, s_t, a_t) \triangleq \mathcal{N}(\mu(w_k, s_t, a_t), 1).$$

Motivation

직관적으로, 초기화도 다르고, 관찰 순서도 다른 앙상블은 처음에는 예측이 다르나,
데이터가 늘어날수록 모델이 동일한 예측으로 수렴하며 불일치가 감소한다.

Exploration

■ Planning in Latent Space

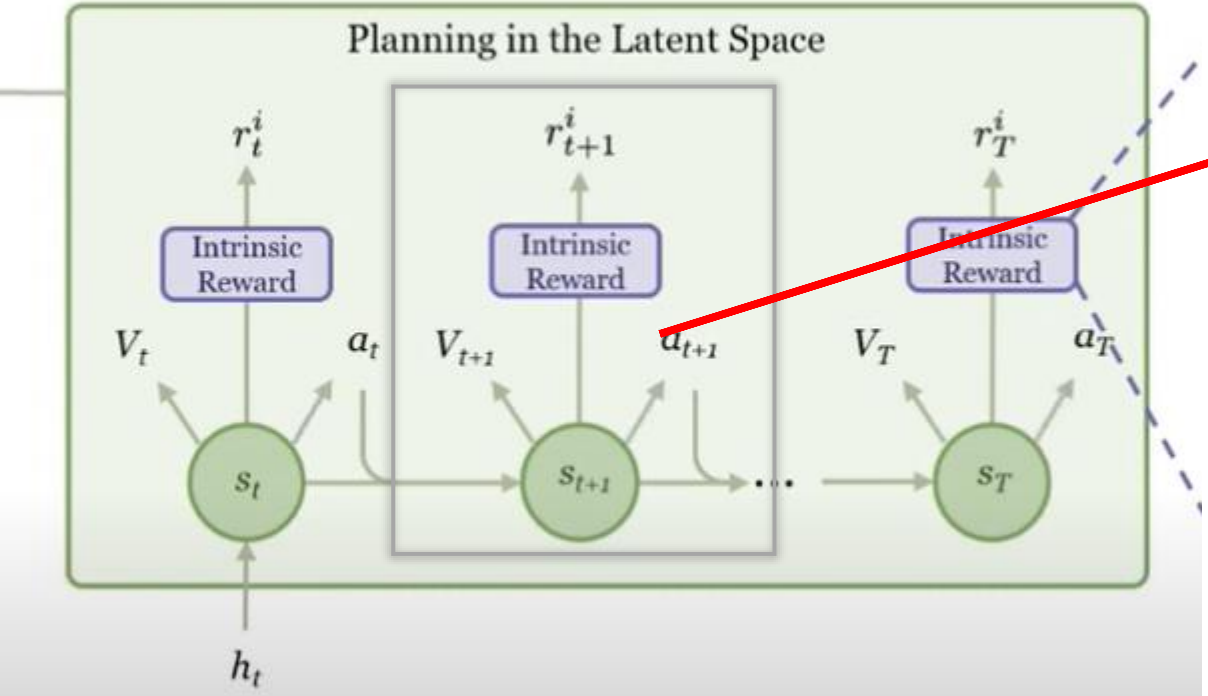
Latent Disgarement를 예측한 평균에 대한 분산으로 정의

$$\begin{aligned} \mathrm{D}(s_t, a_t) &\triangleq \mathrm{Var}\left(\{\mu(w_k, s_t, a_t) \mid k \in [1; K]\}\right) \\ &= \frac{1}{K-1} \sum_k \left(\mu(w_k, s_t, a_t) - \mu'\right)^2, \\ \mu' &\triangleq \frac{1}{K} \sum_k \mu(w_k, s_t, a_t). \end{aligned} \tag{4}$$

Latent Disagreement는
Exploration policy를 훈련시키기 위한 Intrinsic Reward로 사용됨

Exploration

■ Exploration Policy : Train by Dreamer



Dreamer

Exploration

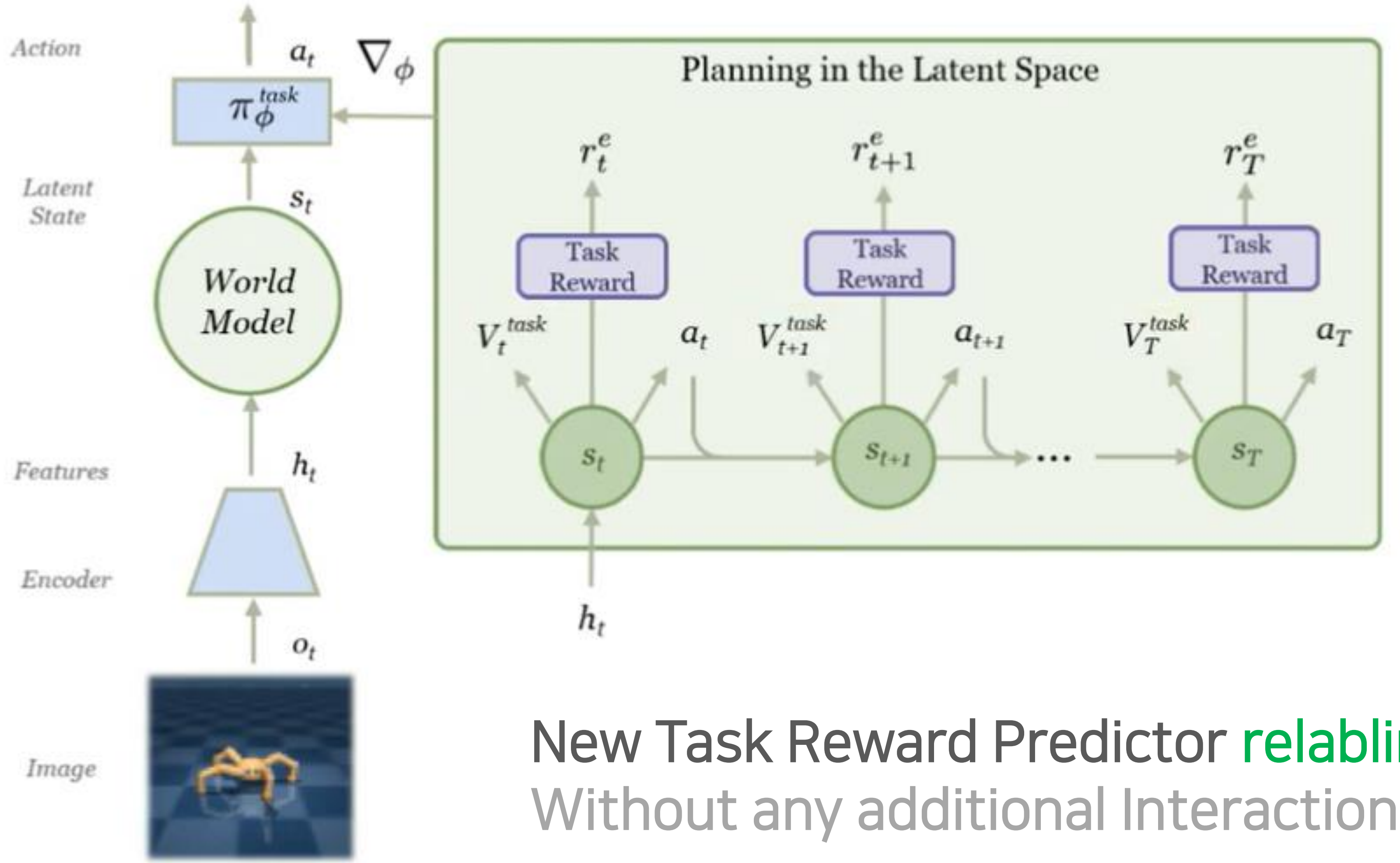
Algorithm 1 Planning to Explore via Latent Disagreement

---

initialize: Dataset D from a few random episodes.
World model M.
Latent disagreement ensemble E.
Exploration actor-critic π_{LD}.
while exploring **do**
    Train M on D.
    Train E on D.
    Train π_{LD} on LD reward in imagination of M.
    Execute π_{LD} in the environment to expand D.
end while
return Task-agnostic D and M.

Solving Task

Solving Task



New Task Reward

New Task Reward Predictor relabling the replay buffer

Without any additional Interaction

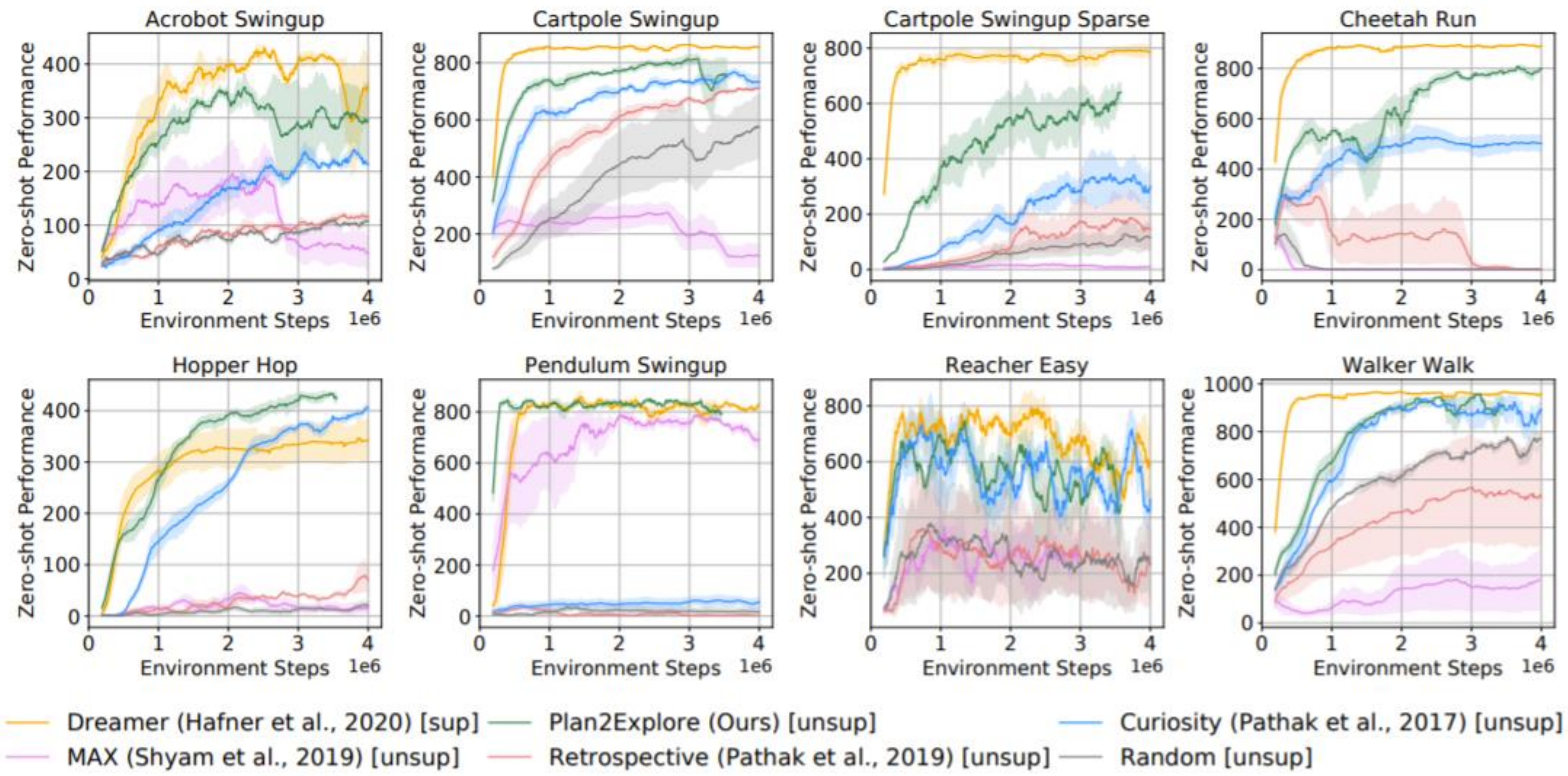
Solving Task

Algorithm 2 Zero and Few-Shot Task Adaptation

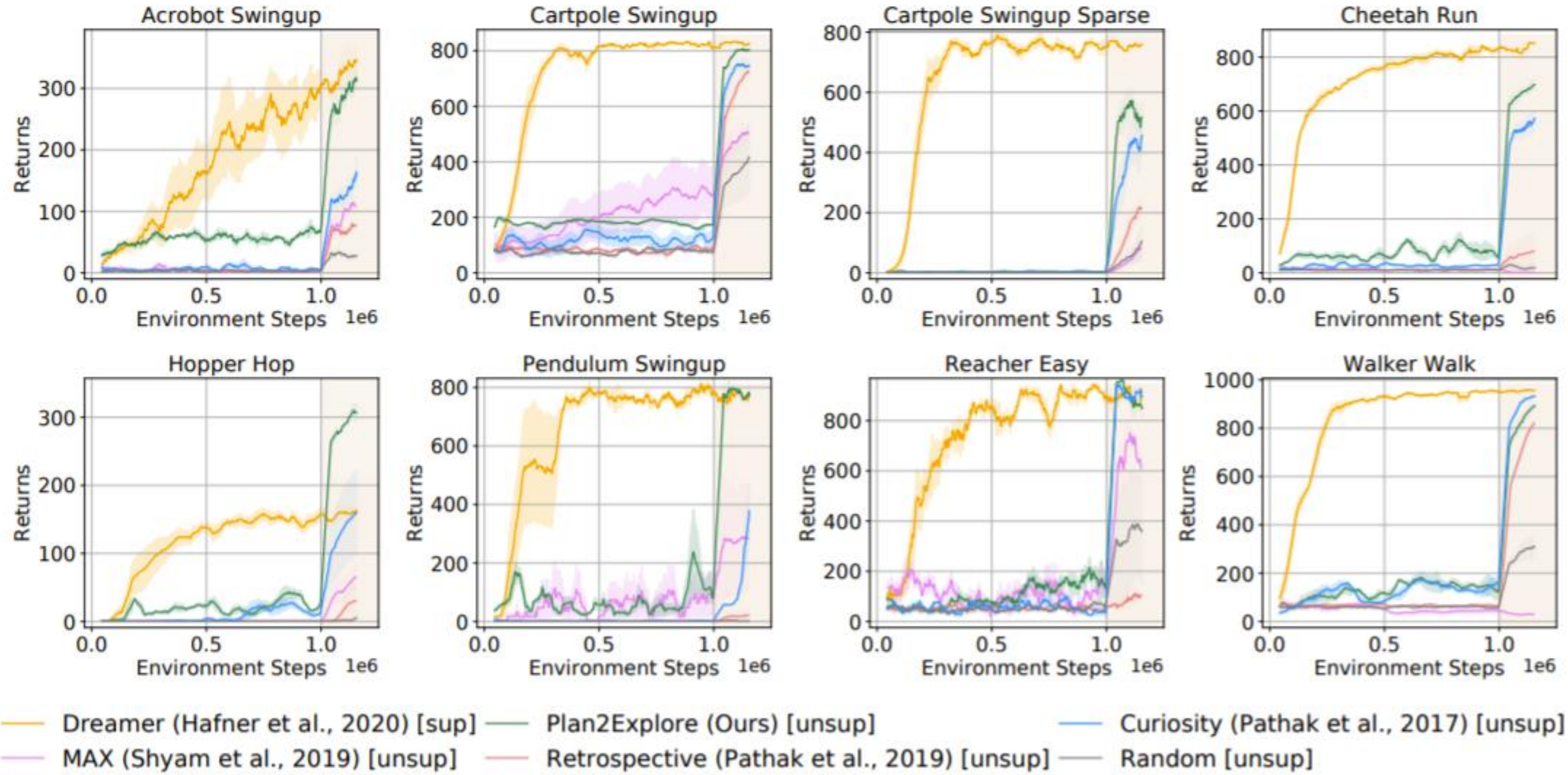
input: World model M.
Dataset D without rewards.
Reward function R.
initialize: Latent-space reward predictor $\hat{\mathrm{R}}$.
Task actor-critic π_{R}.
while adapting **do**
    Distill R into $\hat{\mathrm{R}}$ for sequences in D.
    Train π_{R} on $\hat{\mathrm{R}}$ in imagination of M.
    Execute π_{R} for the task and report performance.
    Optionally, add task-specific episode to D and repeat.
end while
return Task actor-critic π_{R}.

Experiment

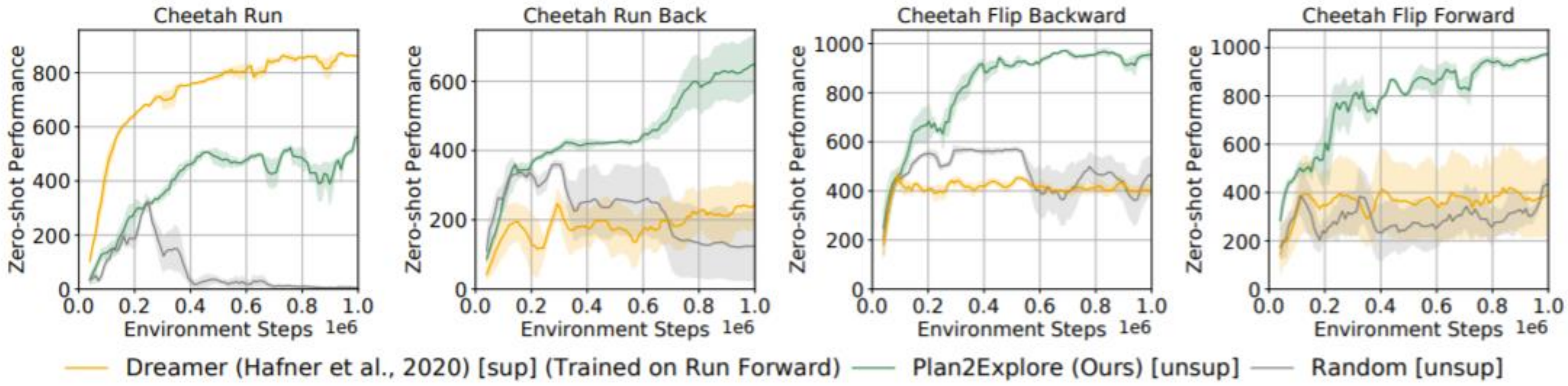
Solving a new-task in zero-shot



Few shot adaptation



Multi Task Performance



왜 Latent Disagreement ?

Expected Information Gain

Latent Disagreement

Maximize Information Gain

$$a_t^* \triangleq \arg\max_{a_t} \mathrm{I}(h_{t+1}; w \mid s_t, a_t).$$

Information Gain

$$\mathrm{I}(h_{t+1}; w \mid s_t, a_t) = \mathrm{H}(h_{t+1} \mid s_t, a_t) - \mathrm{H}(h_{t+1} \mid w, s_t, a_t).$$

$$p(w) \triangleq \frac{1}{K} \sum_k \delta(w - w_k)$$

$$p(h_{t+1} \mid w_k, s_t, a_t) \triangleq \mathcal{N}(h_{t+1} \mid \mu(w_k, s_t, a_t), \sigma^2).$$

$$\mathrm{D}(s_t, a_t) \triangleq \frac{1}{K-1} \sum_k \left(\mu(w_k, s_t, a_t) - \mu'\right)^2,$$

$$\mu' \triangleq \frac{1}{K} \sum_k \mu(w_k, s_t, a_t).$$

End