Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

MIN Faculty
Department of Informatics

Dream to Control

Learning Behaviors by Latent Imagination

University of Hamburg
Faculty of Mathematics, Informatics and Natural Sciences
Department of Informatics

Technical Aspects of Multimodal Systems

07. December 2023

Outline

Motivation

Dreamer

- Learning latent dynamics
- Learning in Latent Space

Dreaming

DreamerV2

DayDreamer

Conclusion

Appendix

Advantages

- No physical robot
- Optimal environment
- Parallel learning
- No supervison

Issues

- Simulation required
- Reduced complexity
- Simulation inaccuracies

1. Off-line training (no simulation)
2. Limited samples
3. High-dim continuous state/action space
4. Safty constraints
5. Partially observable tasks

...

1

[1] Dulac-Arnold, Mankowitz, and Hester 2019.

Dreamer

Published as a conference paper at ICLR 2020

DREAM TO CONTROL: LEARNING BEHAVIORS BY LATENT IMAGINATION

Danijar Hafner *
University of Toronto
Google Brain

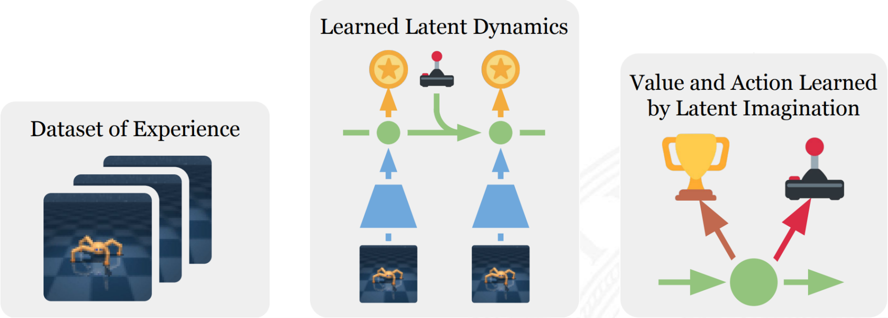
Timothy Lillicrap
DeepMind

Jimmy Ba
University of Toronto

Mohammad Norouzi
Google Brain

[2]

[2]Hafner, Lillicrap, Ba, et al. 2020.


Dataset of Experience
Learned Latent Dynamics
Value and Action Learned by Latent Imagination

Dreamer

Agent Components

- Dynamics learning
- Behavior learning
- Environment interaction

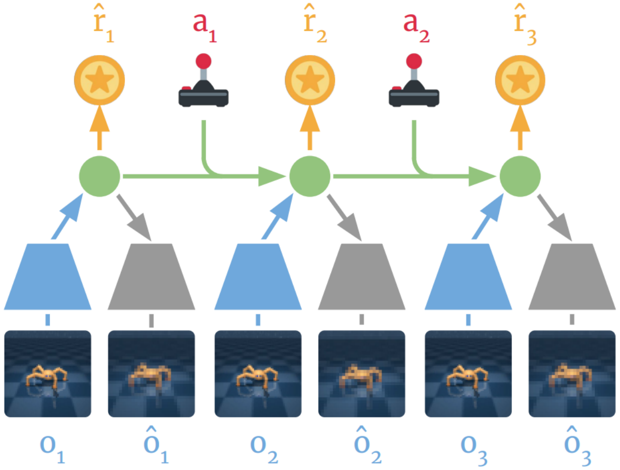
Latent Dynamics Model Components

- Representation model: $p(s_t \mid s_{t-1}, a_{t-1}, o_t)$
- Transition model: $q(s_t \mid s_{t-1}, a_{t-1})$
- Reward model: $q(r_t \mid s_t)$

Environment Interaction Model

- Actor critic
- Action model: $a_\tau \sim q_\phi(a_\tau \mid s_\tau)$
- Value model: $v_\psi(s_\tau) \approx \mathbb{E}_{q(\cdot|s_\tau)}(\sum_{\tau=t}^{t+H} \gamma^{\tau-t} r_\tau)$

Dreamer
Learning latent dynamics



Learning latent dynamics in Dreamer

1. Draw data sequences $\{(a_t, o_t, r_t)\}_{t=k}^{k+L}$ from initial dataset
2. Compute model states $s_t \sim p_\theta(s_t \mid s_{t-1}, a_{t_1}, o_t)$
3. Update neural network parameters θ by representation learning

Reward Prediction

Reconstruction of Image

- Additional observation model: $q(o_t \mid s_t)$ for learning signal

$$\mathcal{J}_{\text{REC}} = \mathbb{E}_p\left(\sum_t (\mathcal{J}_{\text{O}}^t + \mathcal{J}_{\text{R}}^t + \mathcal{J}_{\text{D}}^t)\right) + c$$

$$\mathcal{J}_{\text{O}}^t = \ln q(o_t \mid s_t) \quad \mathcal{J}_{\text{R}}^t = \ln q(r_t \mid s_t)$$

$$\mathcal{J}_{\text{D}}^t = -\beta \text{KL}(p(s_t \mid s_{t-1}, a_{t-1}, o_t) \mid\mid q(s_t \mid s_{t-1}, a_{t-1}))$$

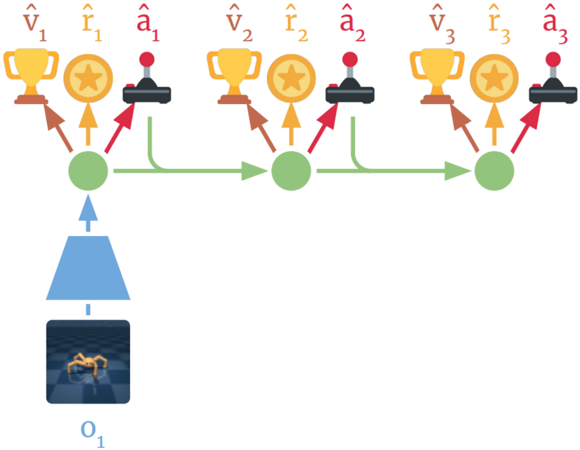
Contrastive Estimation

- Predict state from observation
- State model: $q(s_t \mid o_t)$
- Noise contrastive estimation
  - Averaging over o' = observations of current sequence batch

$$\mathcal{J}_{\text{NCE}} = \mathbb{E}_p \left(\sum_t (\mathcal{J}_{\text{S}}^t + \mathcal{J}_{\text{R}}^t + \mathcal{J}_{\text{D}}^t) \right)$$

$$\mathcal{J}_{\text{S}}^t = \ln q(s_t \mid o_t) - \ln \left(\sum_t q(s_t \mid o') \right)$$

Dreamer
Learning in Latent Space



Imagine future actions, values and rewards

1. Imagine trajectories $\{(s_\tau, a_\tau)\}_{\tau=t}$ from each s_t
2. Predict rewards $\mathbb{E}(q_\theta(r_\tau \mid s_\tau))$ and values $v_\psi(s_\tau)$
3. Compute value estimation $V_\lambda(s_\tau)$
4. Update action model - ϕ and value model parameters ψ

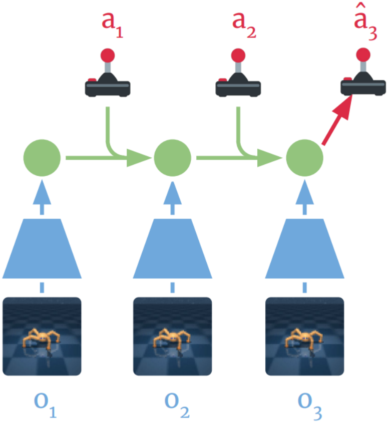
Value Estimation

- Exponetially-weighted average of estimates

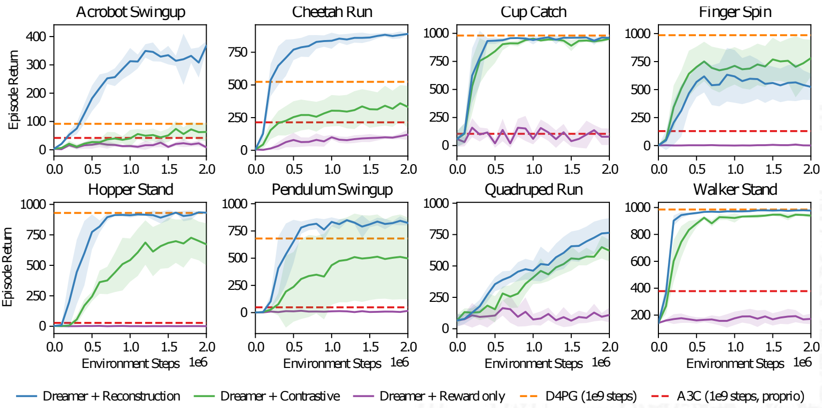
$$\mathrm{V}_\lambda(s_\tau) = (1-\lambda)\sum_{n=1}^{H-1}\lambda^{n-1}\mathrm{V}_\mathrm{N}^n(s_\tau) + \lambda^{H-1}\mathrm{V}_\mathrm{N}^H(s_\tau)$$

$$\mathrm{V}_\mathrm{N}^k(s_\tau) = E_{q_\theta,q_\phi}\left(\sum_{n=\tau}^{h-1}\gamma^{n-\tau}r_n + \gamma^{h-\tau}v_\psi(s_h)\right)$$

with $h = \min(\tau + k, t + H)$

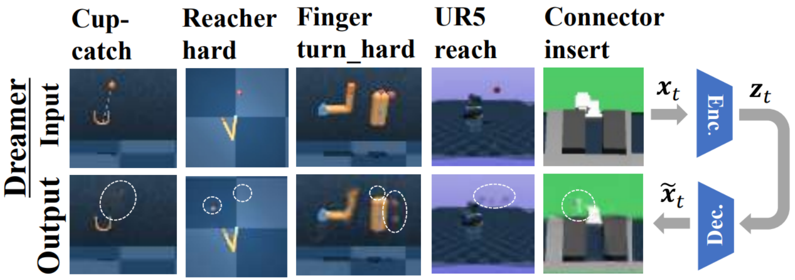


Use trained model to act in environment



Different representation learning objectives in Dreamer

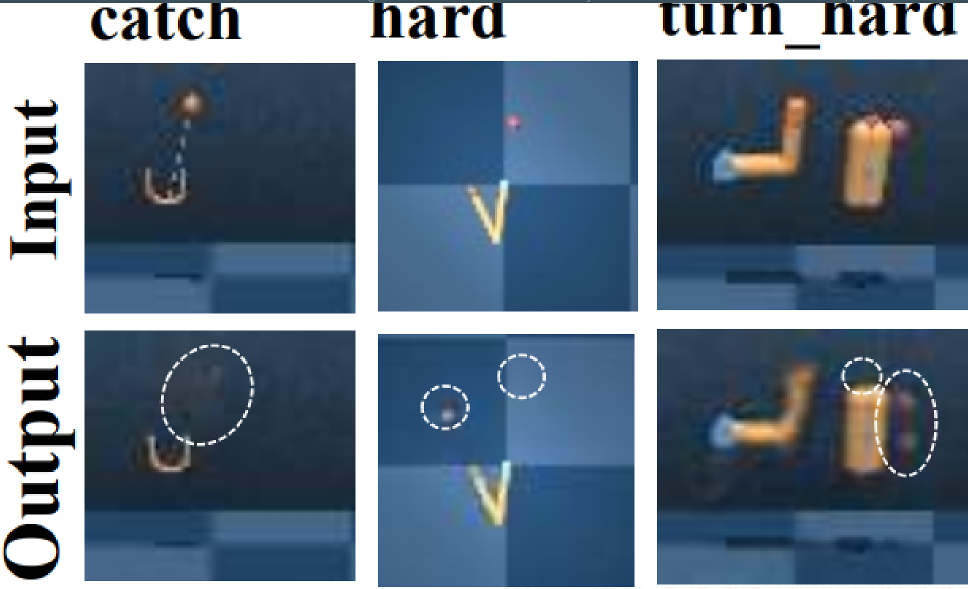
Dreamer
Object Vanishing



Object vanishing[3]

[3] Okada and Taniguchi 2021.

Dreamer
Object Vanishing



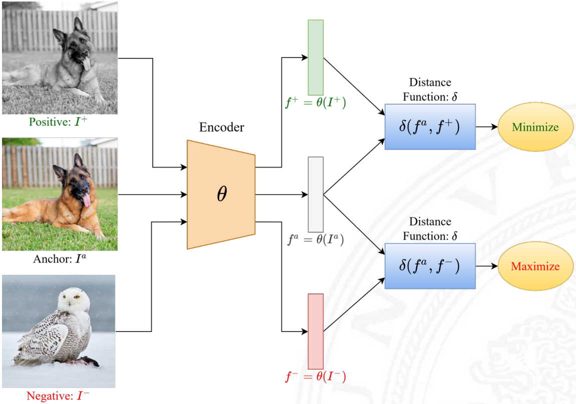
Dreaming: Model-based Reinforcement Learning by Latent Imagination without Reconstruction

Masashi Okada[1,⋆] and Tadahiro Taniguchi[1,2]

4

[4] Okada and Taniguchi 2021.

Dreaming
Contrastive Learning



Contrastive learning[5]

[5] Kundu 2022.

Dreaming
Independent linear Dynamics

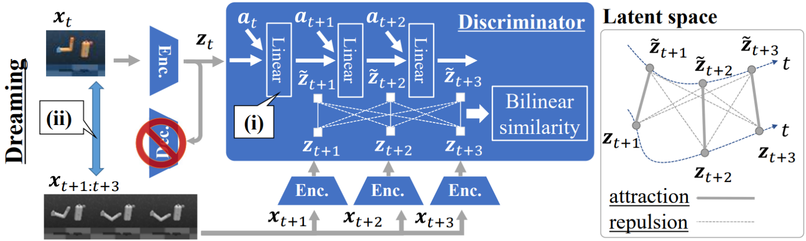
- Multi-step prediction model:
  $\tilde{p}(z_t \mid z_{t-k}, a_{<t}) := \mathbb{E}_{\tilde{p}(z_{t-1}|z_{t-k}, a_{<t-1})}[\tilde{p}(z_t \mid z_{z-1}, a_{t-1})]$
- Hyperparameter k: lantent overshooting
- New objective:

$$\mathcal{J} := \sum_{k=0}^{K} (\mathcal{J}_k^{\text{NCE}} + \mathcal{J}_k^{\text{KL}})$$

$$\mathcal{J}_k^{\text{NCE}} := \mathbb{E}_{\tilde{p}(z_t|z_{t-k}, a_{<t})q(z_{t-k}|\cdot)} \left[\ln p(z_t \mid x_t) - \ln \sum_{x' \in D} p(z_t \mid x') \right]$$

$$\mathcal{J}_k^{\text{KL}} := \mathbb{E}_{p(z_t|z_{t-k}, a_{<t})q(z_{t-k}|x_{\leq t-k}, a_{<t-k})} \hookleftarrow$$
$$[\text{KL}(q(z_{z+1}|x_{\leq t+1}, a_{<t+1}) \mid\mid p(z_{t+1}|z_t, a_t))]$$

- Two independent image preprocessors
- Random Crop $(72, 72) \rightarrow (64, 64)$

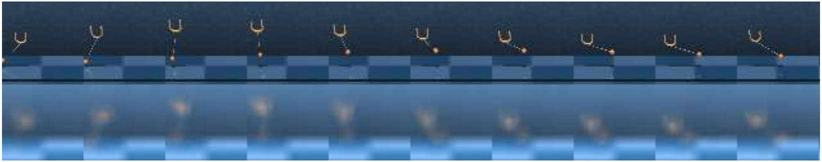


Latent space learning without discriminator[6]

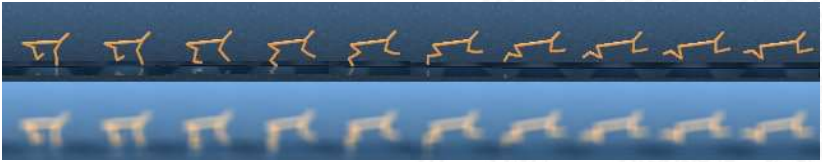
[6] Okada and Taniguchi 2021.

Dreaming

Open-loop video prediction by seperatly trained decoder



Cup-Catch



Cheeta-run

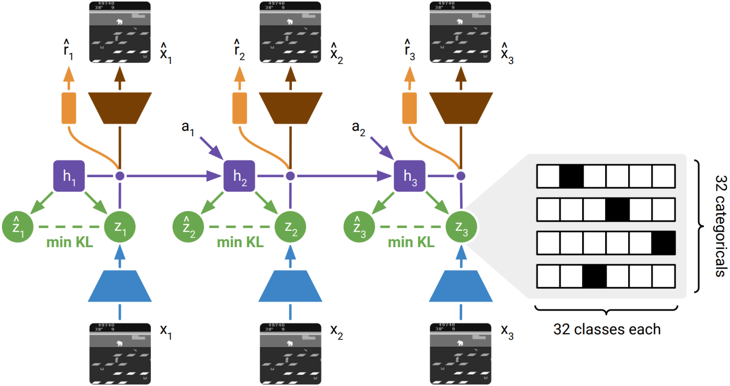
Dreaming
Open-loop video prediction by seperatly trained decoder

DreamerV2

Mastering Atari with Discrete World Models

- Discrete latents
  - Vector of categorical variables
  - Dreamer: diagonal Gaussian
- Balancing terms within KL loss
  - Posterior stochastic state z_t
  - Prior stochastic state $\hat{z}_t$
  - Recurrent model: $h_t = f_\theta(h_{t-1}, z_{t-1}, a_{t-1})$
  - Increase in robustness to novel inputs
- Discount predictor: $\hat{\gamma}_t \sim p_\theta(\hat{\gamma} \mid h_t, z_t)$
  - Estimation of probability of episode ending

DreamerV2



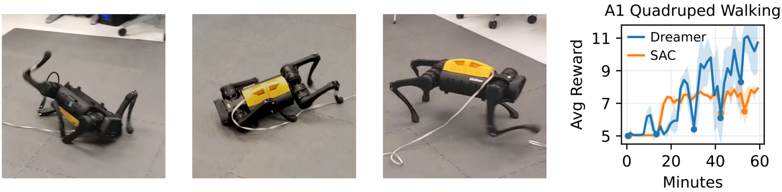
DreamerV2 with discrete latent representation[7]

[7] Hafner, Lillicrap, Norouzi, et al. 2022.

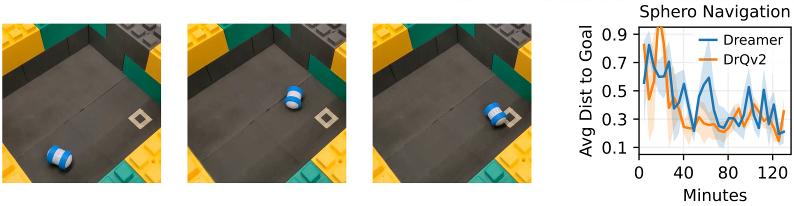
DayDreamer

Online Reinforcement Learning in Real World

- No simulations or demonstrations
- Learner and Actor thead
- Parallel training
- Sensor fusion in encoder
- Different gradient estimators for continuous/discrete tasks
- Identical hyperparameters for all experiments

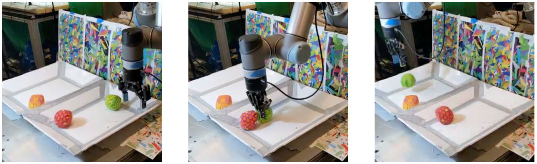


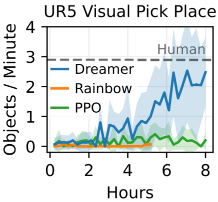
Quadruped walking after 1 hour Wu et al. 2023



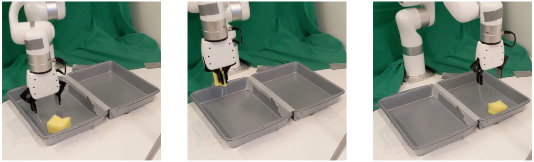
Sphero navigation in 2 hours

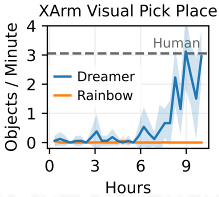
DayDreamer
Experiments





Visual pick and place with multiple objects on UR5 (2.5 objects/min after 8 h)





Visual pick and place with XArm

DayDreamer
Experiment Video

https://danijar.com/project/daydreamer/

Conclusion

- Data/computation time-efficient actor critic method
  - Dreamer
  - Dreaming
  - DreamerV2
  - (DreamingV2)
- human-level performance
- Image/multi-modal input
- Successfully tested on real robots
- Object-vanishing/early over-fitting
- Hardware wear

Thank you for your attention.

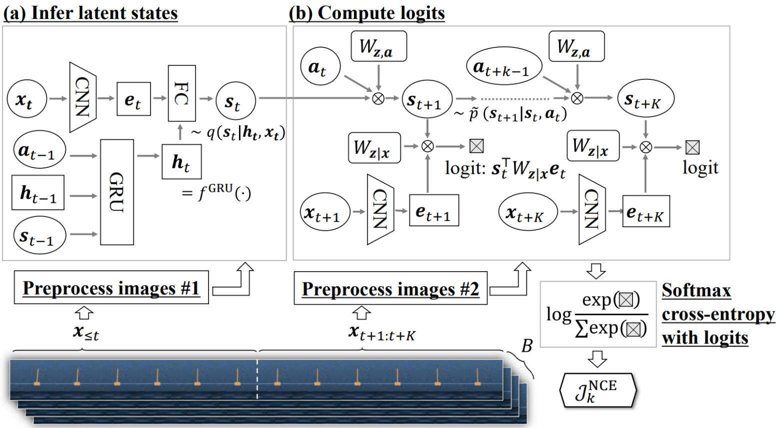
Are there any questions ?

Kullback-Leibler Divergence[8]

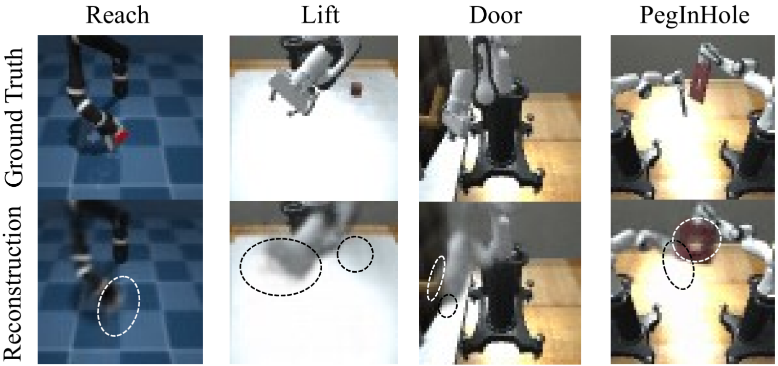
- $\mathrm{KL}(P \parallel Q)$
- P: "true" distribution of data
- Q: approximation of P
- Relative entropy of P with respect to Q
- Amount of information lost when Q is used instead of P

[8] contributors 2023.

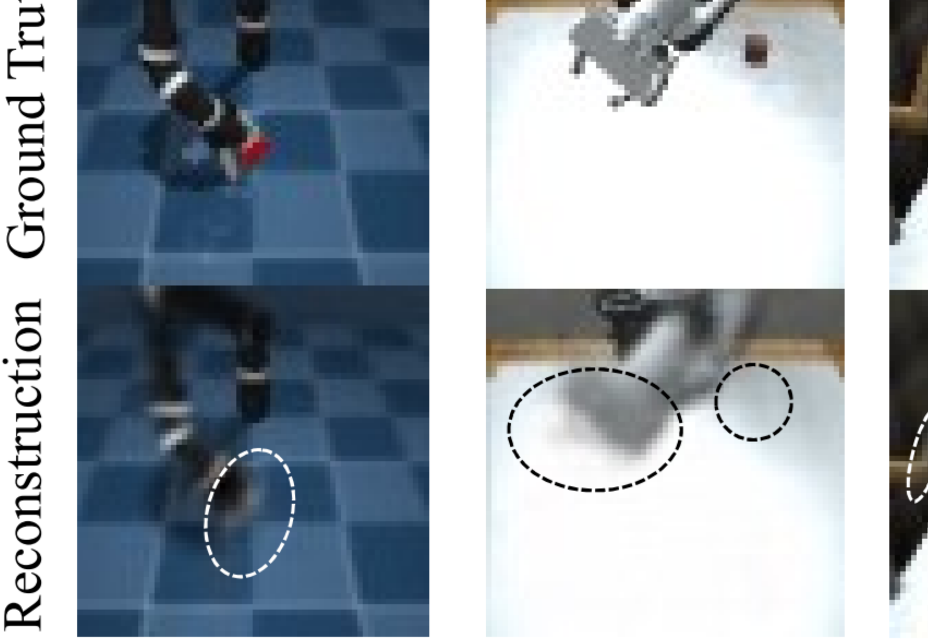
Dreaming-Details



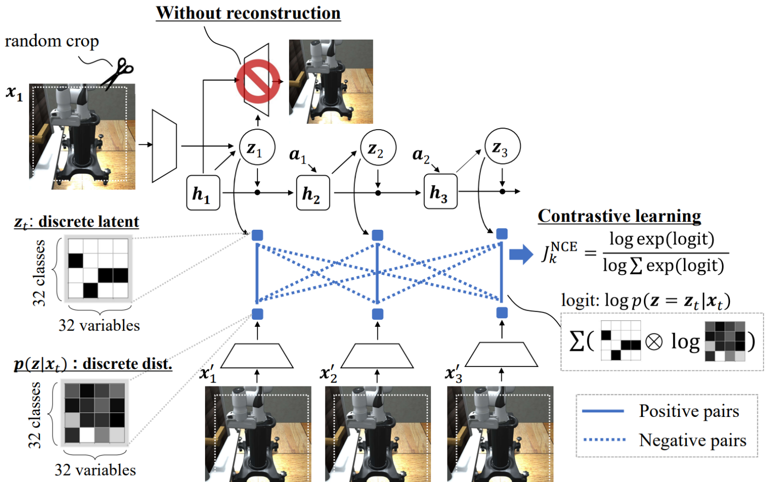
Dreaming(V1) latent representation learning Okada and Taniguchi 2021



Object vanishing in DreamerV2 Okada and Taniguchi 2022

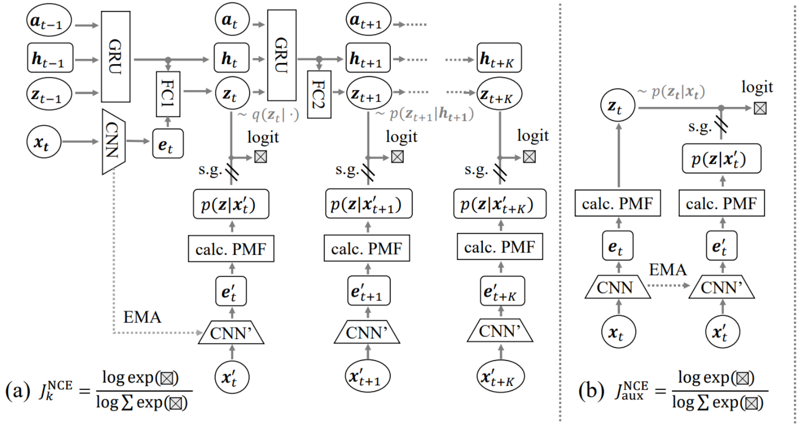

Ground Tru
Reconstruction

DreamerV2



DreamingV2 Okada and Taniguchi 2022

DreamingV2-Details



DreamingV2 latent representation learning Okada and Taniguchi 2022

Comparison of Dreamer and Dreaming

	DreamingV2	DreamerV2	Dreaming	Dreamer
Discrete latent	✓	✓		
Reconstruction free	✓		✓	
3D Robot-arm				
UR5-reach	**776**±194	704±222	752±1178	701±223
Reach-duplo	**199**±43	149±62	145±61	5±11
Lift	**327**±150	165±126	174±107	134±46
Door	**383**±143	190±126	319±173	154±32
PegInHole	**436**±26	376±59	353±50	354±47
2D Manipulation				
Reacher-hard	598±447	175±340	**743**±346	247±392
Finger-turn-hard	484±434	600±417	**858**±210	533±426
Reacher-easy	924±210	923±215	**947**±100	658±429
Finger-trun-easy	434±469	498±469	**842**±286	665±430
2D Locomotion				
Cheetah-run	768±24	**811**±75	542±132	776±120
Walker-walk	857±115	**951**±28	518±76	906±70

Scores on different robot tasks Okada and Taniguchi 2022

contributors, Wikipedia (Dec. 4, 2023). *Kullback–Leibler divergence*. In: *Wikipedia*. Page Version ID: 1188277594. URL: https://en.wikipedia.org/w/index.php?title=Kullback%E2%80%93Leibler_divergence&oldid=1188277594 (visited on 12/06/2023).

Dulac-Arnold, Gabriel, Daniel Mankowitz, and Todd Hester (Apr. 29, 2019). *Challenges of Real-World Reinforcement Learning*. arXiv: 1904.12901[cs,stat]. URL: http://arxiv.org/abs/1904.12901 (visited on 11/29/2023).

Hafner, Danijar, Timothy Lillicrap, Jimmy Ba, et al. (Mar. 17, 2020). *Dream to Control: Learning Behaviors by Latent Imagination*. arXiv: 1912.01603[cs]. URL: http://arxiv.org/abs/1912.01603 (visited on 11/26/2023).

Hafner, Danijar, Timothy Lillicrap, Mohammad Norouzi, et al. (Feb. 12, 2022). *Mastering Atari with Discrete World Models*. arXiv: 2010.02193[cs,stat]. URL: http://arxiv.org/abs/2010.02193 (visited on 11/26/2023).

Sources (cont.)

Kundu, Rohit (May 22, 2022). *The Beginner's Guide to Contrastive Learning*. V7. URL: https://www.v7labs.com/blog/contrastive-learning-guide,%20https://www.v7labs.com/blog/contrastive-learning-guide (visited on 11/29/2023).

Okada, Masashi and Tadahiro Taniguchi (Mar. 11, 2021). *Dreaming: Model-based Reinforcement Learning by Latent Imagination without Reconstruction*. arXiv: 2007.14535[cs,eess,stat]. URL: http://arxiv.org/abs/2007.14535 (visited on 11/26/2023).

– (Mar. 1, 2022). *DreamingV2: Reinforcement Learning with Discrete World Models without Reconstruction*. arXiv: 2203.00494[cs,eess]. URL: http://arxiv.org/abs/2203.00494 (visited on 11/26/2023).

Sources (cont.)

Wu, Philipp et al. (Mar. 6, 2023). "DayDreamer: World Models for Physical Robot Learning". In: *Proceedings of The 6th Conference on Robot Learning*. Conference on Robot Learning. ISSN: 2640-3498. PMLR, pp. 2226–2240. URL: https://proceedings.mlr.press/v205/wu23c.html (visited on 11/26/2023).