NEURIPS 2026

NASDAQ

Normalized Observation Space Dynamics-Augmented Q-Learning

Observation normalization enables a unified observation-predictive RL framework to match or surpass model-based and self-predictive RL across diverse domains, while requiring significantly less training wall-time.

Xinwei Liu[1] **Junyuan Liang**[1] **Zicong Hong**[2] **Jianting Zhang**[3] **Wuhui Chen**[1]

[1] Sun Yat-sen University [2] EPFL [3] Purdue University

Code: github.com/sumwailiu/NASDAQ

MOTIVATION

Sample efficiency, without a world model

Model-free RL

TD3, DrQ-v2

Sample-inefficient.

Model-based RL

DreamerV3, TD-MPC2

Improves sample efficiency and performance — **at substantial algorithmic and computational cost.**

Self-predictive RL

TD7, MR.Q

Learns representations by predicting latent embeddings of future observations. **Promising but complex.**

Observation-predictive RL

OFENet

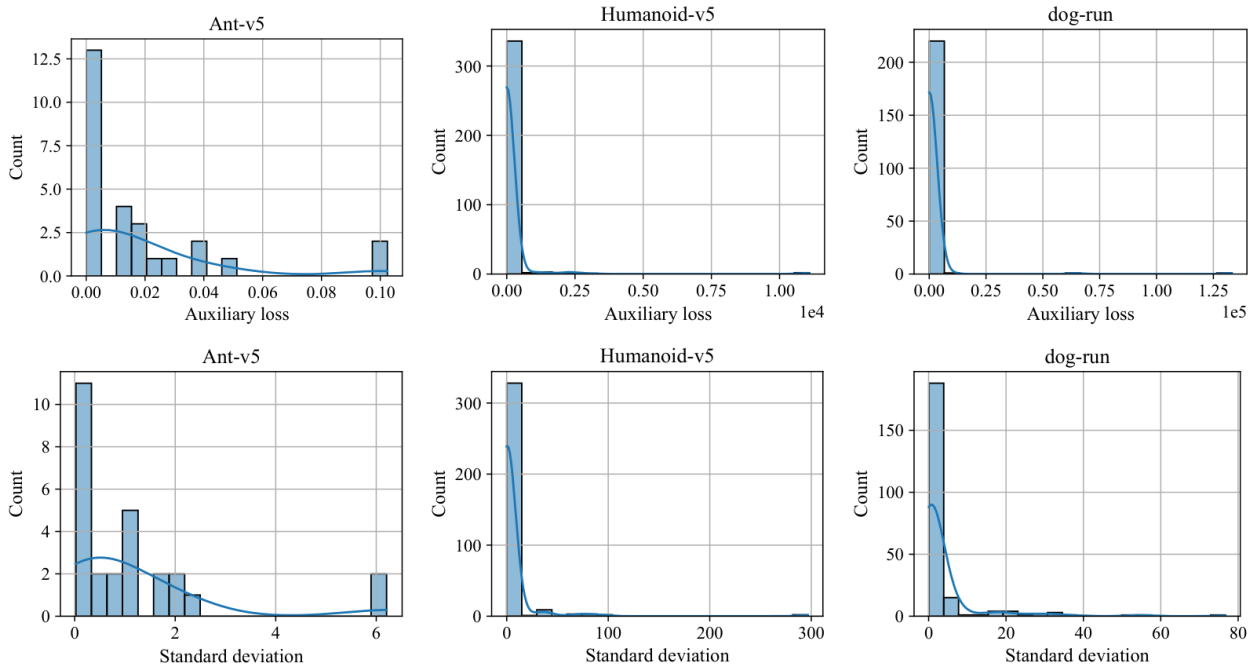
Keep model-free RL, add one auxiliary task: predict the next observation. Limited extra computation, but **still struggles on hard low-dimensional tasks.**

Q. What is the bottleneck in observation-predictive RL — and can it match model-based and self-predictive RL while staying efficient?

Our answer: an unbalanced reconstruction loss across observation dimensions.

DIAGNOSIS

The bottleneck: a skewed reconstruction loss



OFENet+TD3, single run: per-dimension reconstruction loss (top row) and per-dimension standard deviation (bottom row).

The loss profile mirrors the input's value range.

Task	Aux loss	Return
Ant-v5	0.015	8156
Walker2d-v5	0.054	6009
Humanoid-v5	**139**	**6063**
dog-run	**272**	**66**

OFENet+TD3, end of training, 5 seeds; MR.Q gets 300 on dog-run.

- The two tasks with by far the largest loss are the two where it performs worst.
- Per-dimension loss tracks per-dimension **standard deviation** — large-range dimensions dominate the gradient.
- So the learned representations **neglect dimensions with small ranges**.

The bottleneck is an unbalanced reconstruction loss across dimensions.

Fix: normalize the observation space

Shift-Adaptive Robust Observation Normalization (SARON)

$$m \leftarrow \beta\, m + (1-\beta)\,\mu, \qquad s \leftarrow \beta\, s + (1-\beta)\,\nu$$

$$\tilde{o} = \mathrm{clip}\left(\frac{o-\hat{m}}{\hat{\sigma}},\ -O,\ O\right), \qquad \hat{\sigma} = \sqrt{\hat{s} - \hat{m}^2}$$

- Episode-wise mean and second moment, averaged with an EMA and bias-corrected.
- **Why not global statistics?** Online RL shifts the distribution as the policy improves, and exploration inflates the variance.
- **Robustness:** statistics probably induced by exploration are treated as outliers and excluded by an interquartile-range test over recent returns.

NORMALIZATION ALONE, ON OFENET+TD3

Gym · mean over 5 tasks

1.48 → **1.64**

DMC · IQM over 28 tasks

600 → **656**

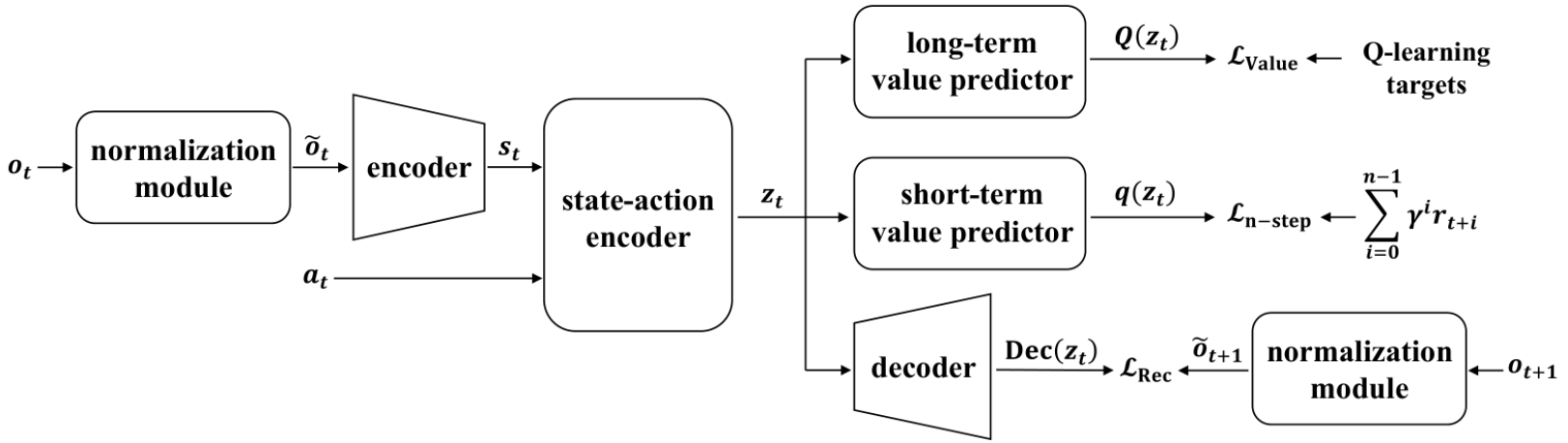
Beats running normalization (SB3) on 9 of 12 metrics.

… and it buys something bigger

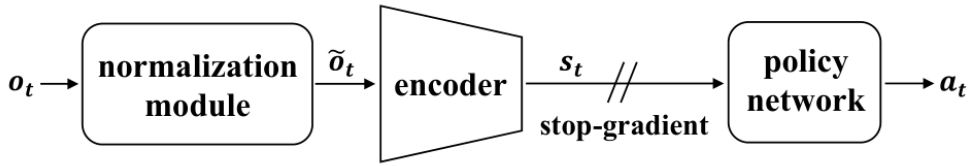
Once observations live in a normalized space, **dynamics prediction is well defined for pixels too** — one predictor for physical states and images.

Normalizing fixes the loss imbalance and unifies low- and high-dimensional inputs.

NASDAQ: one framework, states or pixels



(a) Value network with two auxiliary prediction heads



(b) Policy network with detached inputs

01 Q · long-term value
TD3-style n-step target, Huber loss

02 q · short-term value
n-step return, two-hot cross-entropy

03 Dec · next observation
predicts the next normalized observation

Low-dimensional control: best mean and IQM on both

Gym (5 tasks)	Mean	Median	IQM
NASDAQ (ours)	**1.79**	**1.26**	**1.50**
MR.Q	1.50	1.12	1.36
TD7	1.16	**1.30**	1.14
OFENet+TD3	1.48	1.24	1.44
TD-MPC2	0.36	0.20	0.20
DreamerV3	0.66	0.62	0.68
TDx (ours, no aux.)	1.55	1.22	1.30

Scores normalized by a deep TD3 variant.

DMC proprioceptive (28)	Mean	Median	IQM
NASDAQ (ours)	**805**	**927**	**898**
MR.Q	791	918	886
TD-MPC2	783	896	868
TD7	660	810	726
OFENet+TD3	567	696	600
DreamerV3	530	700	577
TDx (ours, no aux.)	744	839	834

Mean / median / IQM over 5 seeds.

- **+19% mean reward on Gym over MR.Q, and higher than MR.Q on all six aggregate metrics.**
- **A by-product — TDx:** remove the auxiliary tasks and the same framework still performs strongly as a plain model-free baseline.

Pixels: best on DMC visual, on par with MR.Q on Atari

DMC visual (28 tasks)	Mean	Median	IQM
NASDAQ (ours)	**606**	**836**	**701**
MR.Q	598	809	686
DrQ-v2	510	626	545
TD-MPC2	492	572	501
DreamerV3	463	493	452

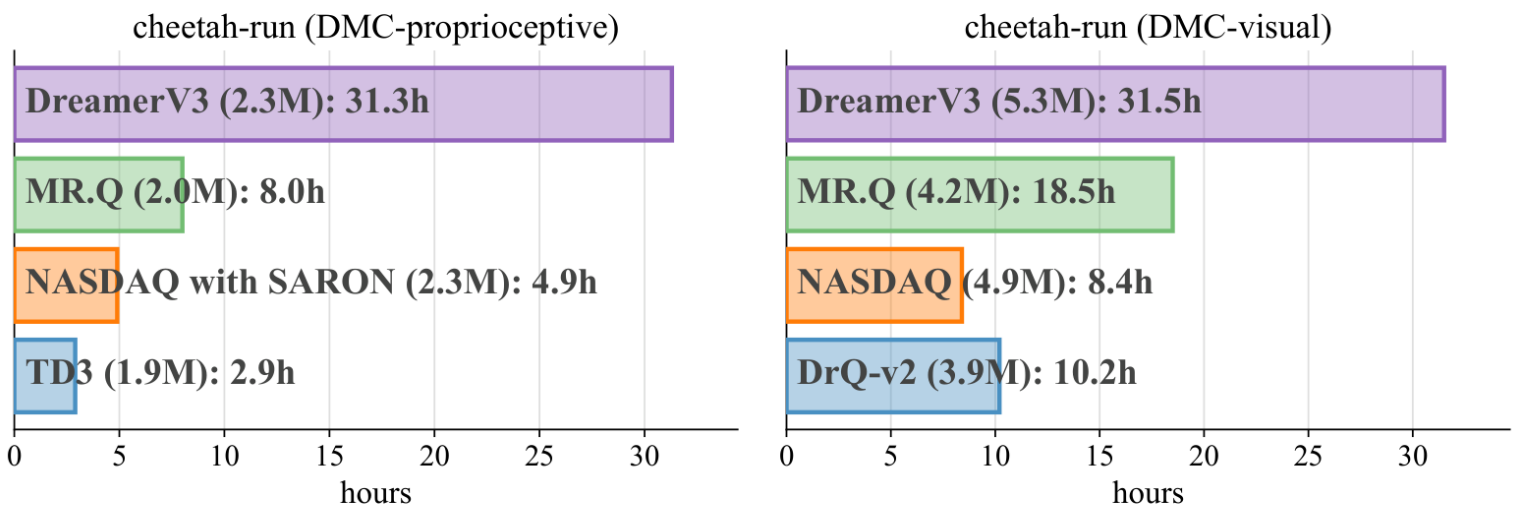
Best mean, median and IQM on DMC visual.

Atari100k (26 games)	Mean	Median	IQM
NASDAQ (ours)	**0.93**	**0.34**	**0.42**
MR.Q	0.91	0.40	0.41
SPR	0.65	0.42	0.43
Rainbow	0.54	0.29	0.35
DreamerV3	**1.25**	**0.49**	**0.54**

Human-normalized. DreamerV3 uses ≈200M parameters — about 40× our 4.9M.

- **One framework, two normalizers:** SARON for physical states, min-max normalization for pixel inputs.
- **Competitive or superior on both continuous-control benchmarks; on Atari100k, a small gap to a 40× larger model.**

Competitive performance at a fraction of the wall-time



4.9 h
vs 8.0 h for MR.Q
state-based

8.4 h
faster than DrQ-v2 (10.2 h)
pixel-based

31.3 h
DreamerV3, same task
world model

cheetah-run, single RTX 4090 — hours to train one task.

- **Where the efficiency comes from:** minor modifications to model-free RL, plus 17% of the parameters in an encoder reused across objectives (DrQ-v2: 0.7%).

TAKEAWAYS

Conclusions

1

The bottleneck is an unbalanced reconstruction loss.

Large-range observation dimensions dominate the gradient; dimensions with small ranges are neglected.

2

Normalize the observation space.

SARON balances the loss across dimensions, and one predictor then serves states and pixels alike.

3

Competitive or superior results, at a fraction of the wall-time.

On 87 environments from four benchmarks, with significantly less training wall-time.

Our hope: **inspire further research on observation-predictive and model-free methods**

Code: **github.com/sumwailiu/NASDAQ**

Thank you!