iQRL: Implicitly Quantized Representations for Sample-Efficient Reinforcement Learning

Aidan Scannell, Kalle Kujanpää, Yi Zhao, Mohammadreza Nakhaei, Arno Solin, Joni Pajarinen

Aidan Scannell
Finnish Center for Artificial Intelligence (FCAI)
Aalto University

Project website

FCAI

fcai.fi

Background

Representation learning for RL

Encoder $z_t = e_\theta(o_t)$

Dynamics $\hat{z}_{t+1} = z_t + d_\phi(z_t, a_t)$

Reward $\hat{r}_{t+1} = r_\phi(z_t, a_t)$

Critic $q_t = Q_\psi(z_t, a_t)$

Policy $a_t \sim \pi_\eta(a_t \mid z_t)$

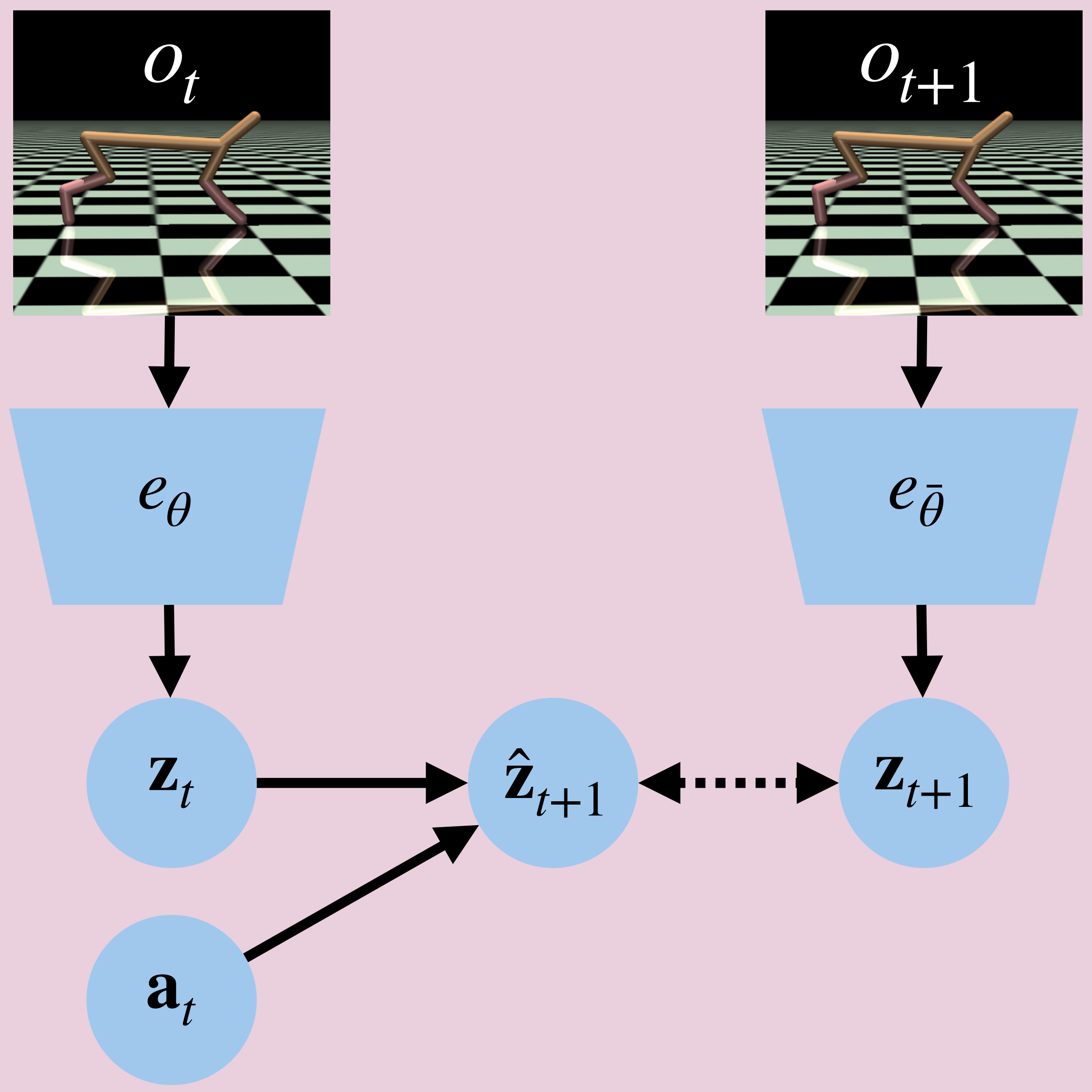
Latent-state consistency loss (representation learning) EMA

$$\arg\min_{\theta,\phi} \sum_{h=t}^{t+H} \gamma^h \boxed{\left(\frac{z_h + d_\phi(e_\theta(o_h), a_h)}{\|z_h + d_\phi(e_\theta(o_h), a_h)\|_2}\right)^\top \left(\frac{\boxed{e_{\bar{\theta}}}(o_{h+1})}{\boxed{e_{\bar{\theta}}}(o_{h+1})}\right)} + \boxed{\left\| r_\phi(e_\theta(o_h), a_h) - r_{h+1} \right\|_2^2}$$

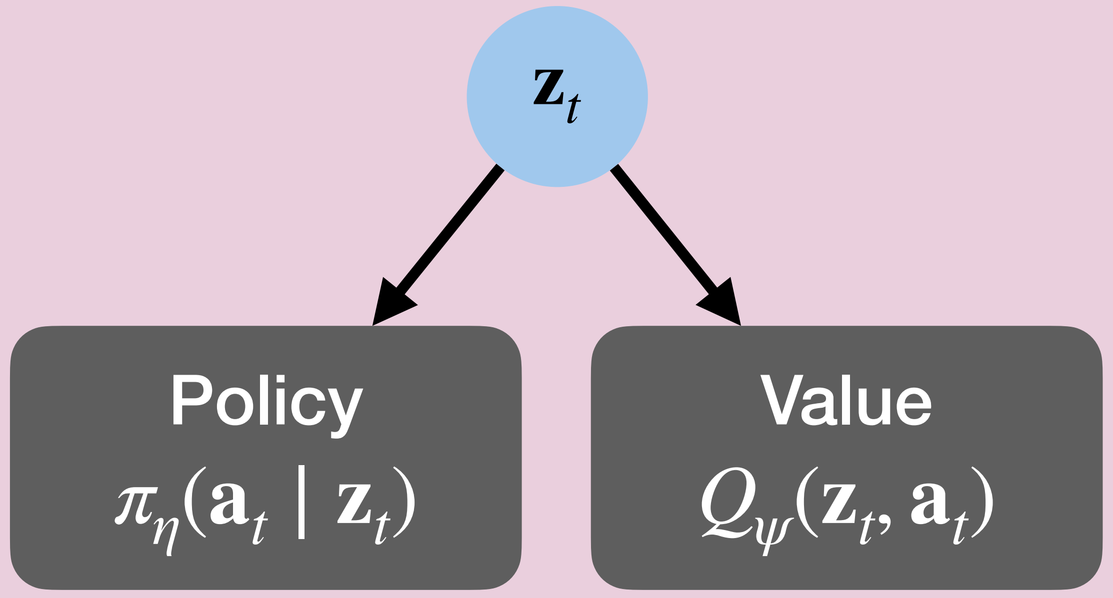
Latent-state consistency with cosine similarity

Representation is task specific!

1. Learn representation



2. Latent actor-critic



Zhao et al. (2023). Simplified Temporal Consistency Reinforcement Learning. ICML.

Background

Task-agnostic representations for RL

Temporal Consistency RL (TCRL)

Remove reward prediction?

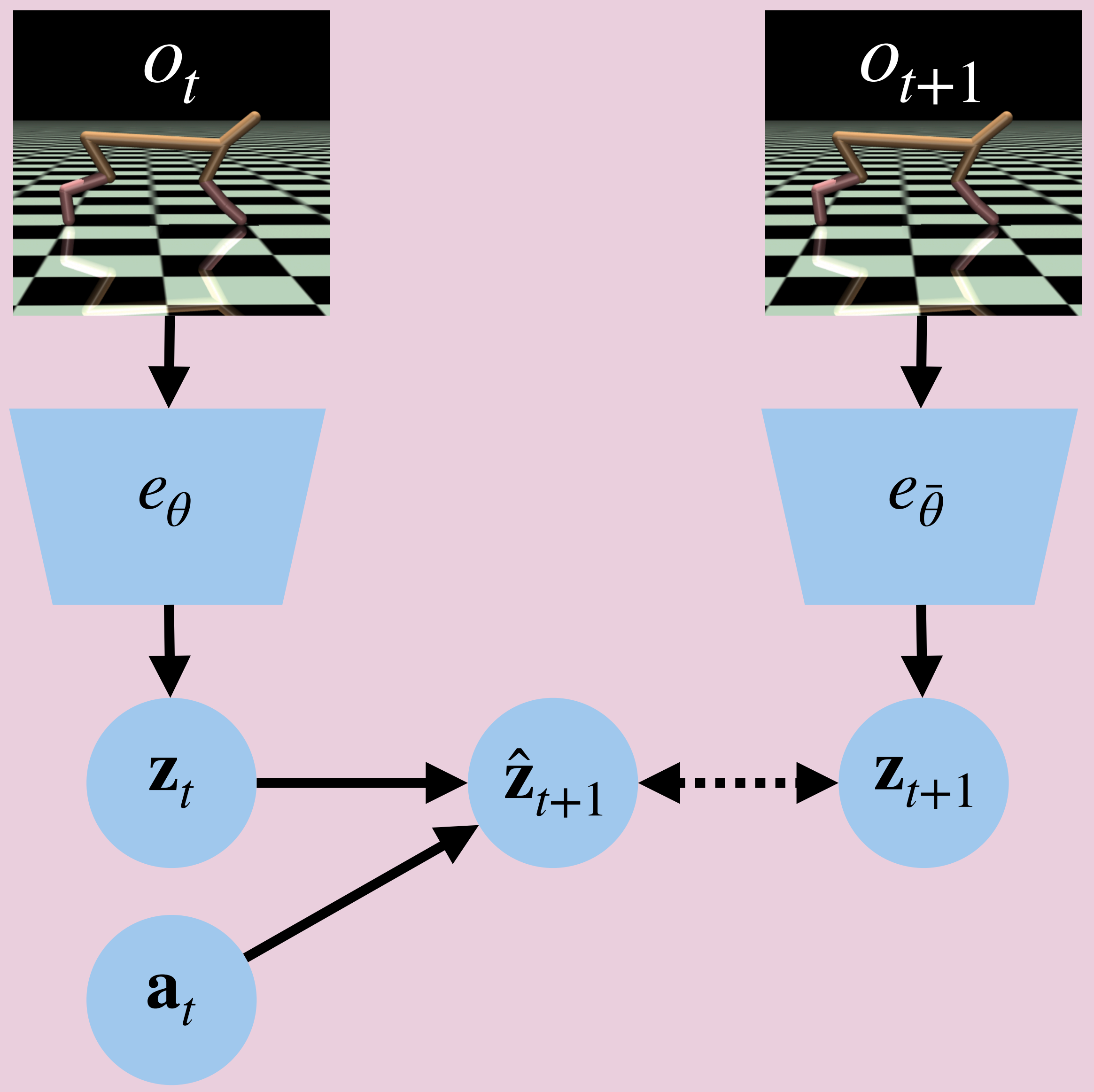
$$\arg\min_{\theta,\phi} \sum_{h=t}^{t+H} \gamma^h \left(\frac{z_h + d_\phi(e_\theta(o_h), a_h)}{\| z_h + d_\phi(e_\theta(o_h), a_h) \|_2} \right)^\top \left(\frac{e_{\bar{\theta}}(o_{h+1})}{\| e_{\bar{\theta}}(o_{h+1}) \|_2} \right) + \boxed{\left\| r_\phi(e_\theta(o_h), a_h) - r_{h+1} \right\|_2^2}$$

Representation is task specific!

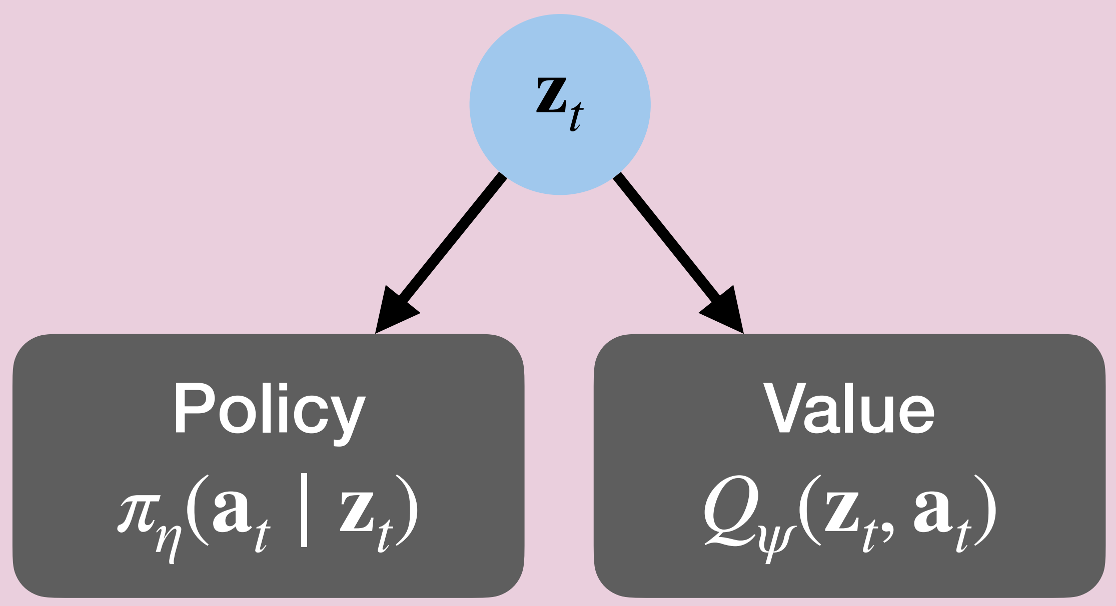
But representation collapse…

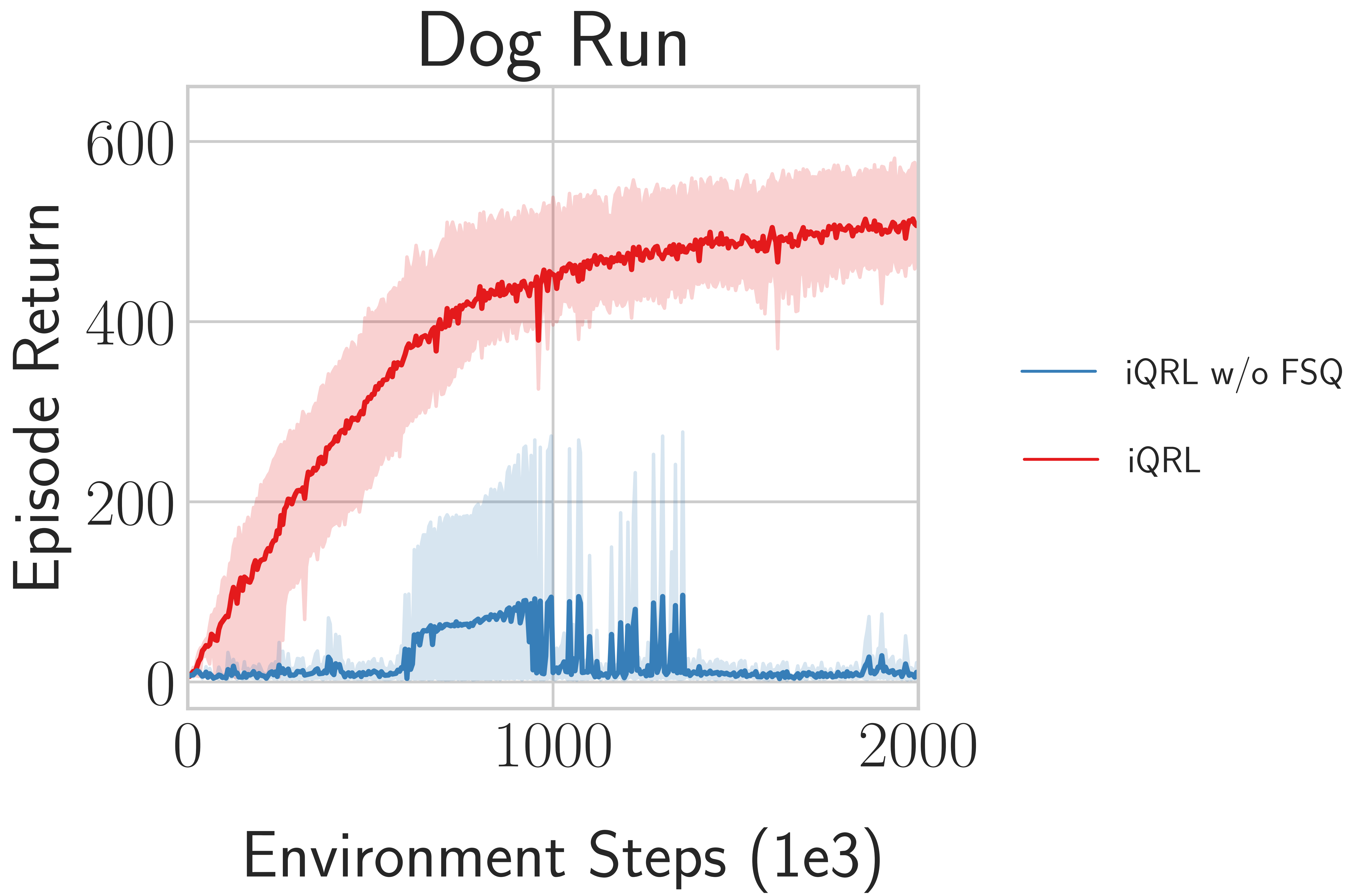
$$e_\theta(o) = \text{const} \quad \forall o \in \mathcal{O}$$

1. Learn representation



2. Latent actor-critic




Dog Run
Episode Return
600
400
200
0
0
1000
2000
Environment Steps (1e3)
iQRL w/o FSQ
iQRL

iQRL

Model-free RL in latent space

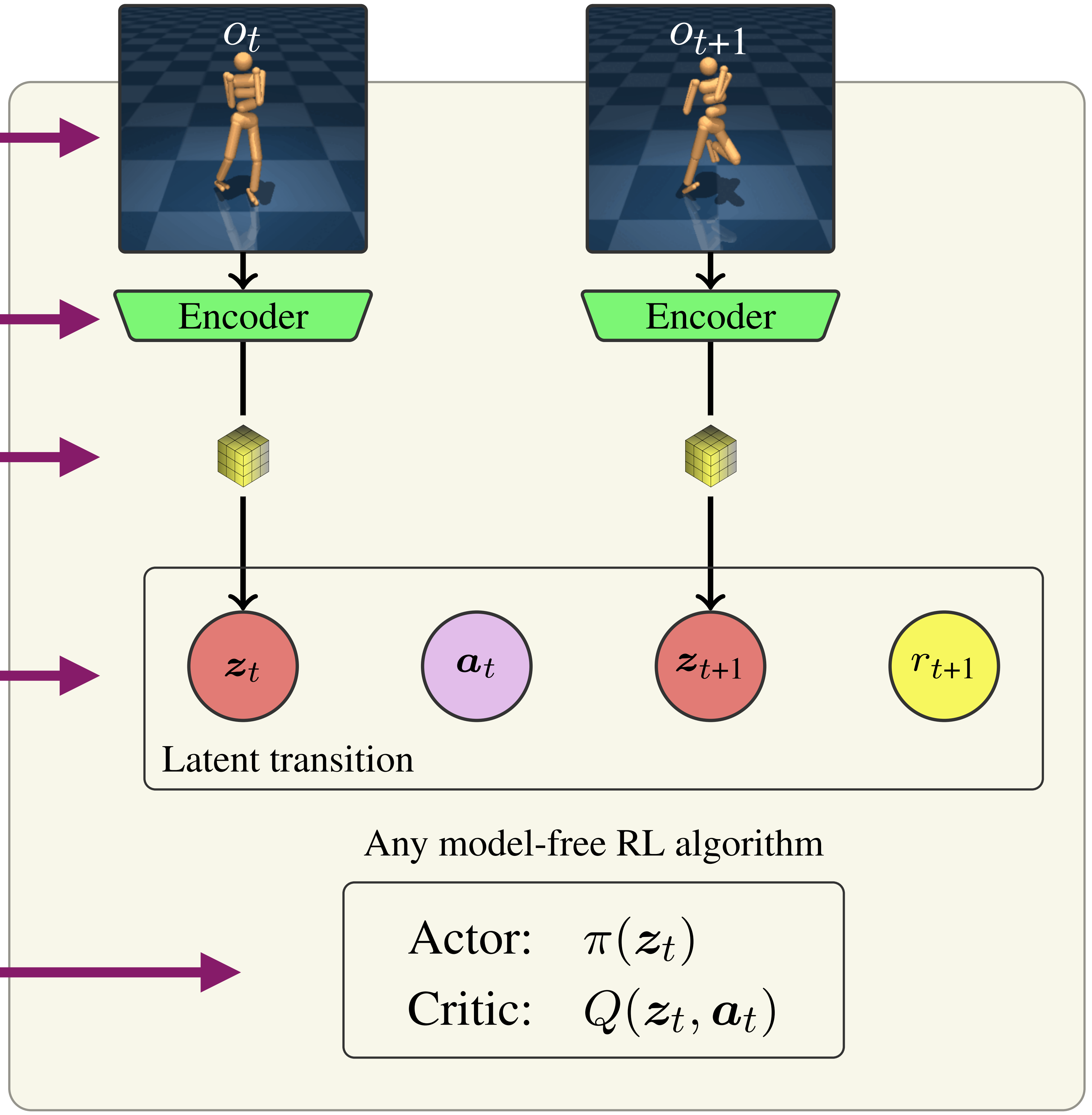
Observation

Encoder

Quantization

Latent transition

Model-free RL
in latent space

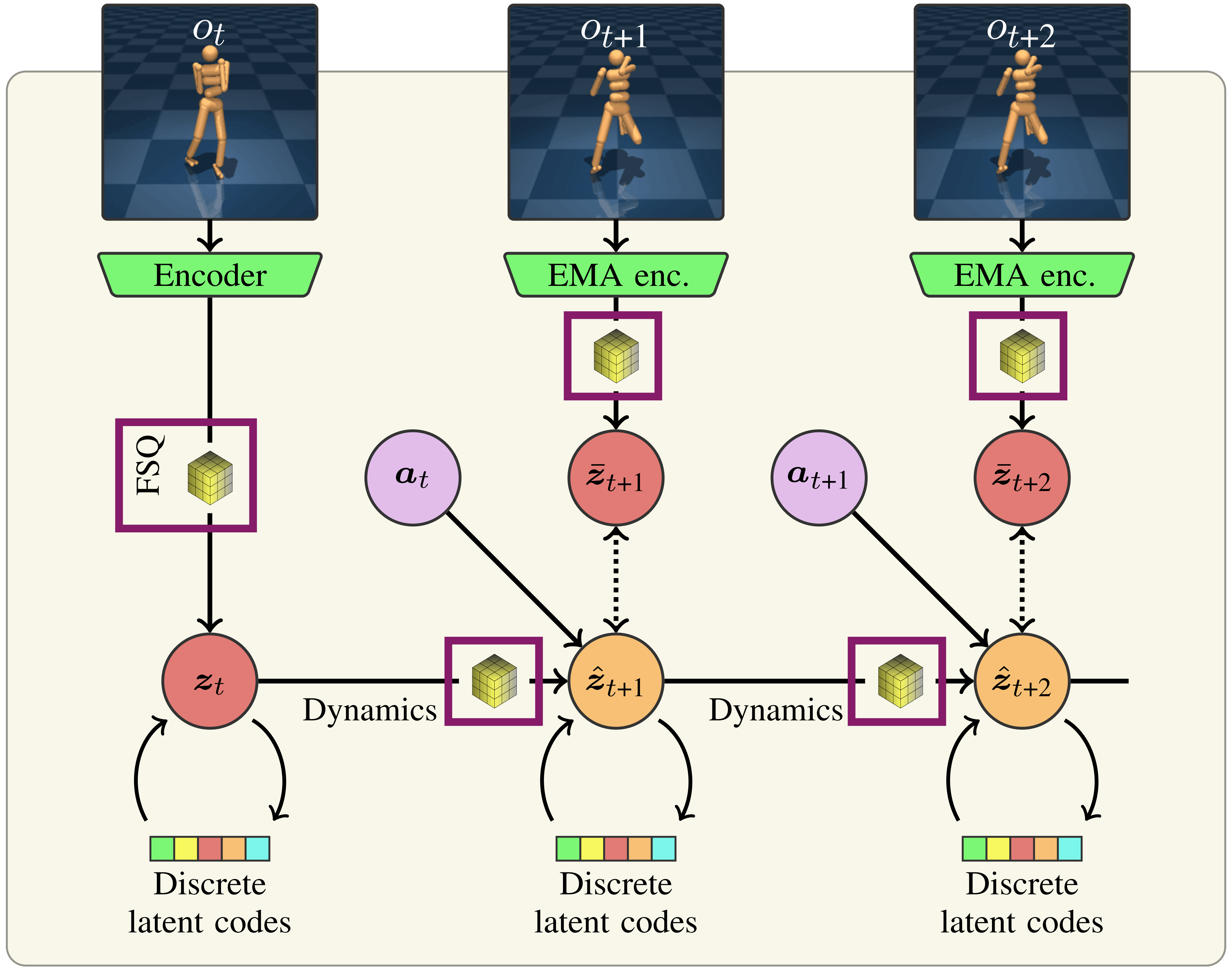


FCAI

fcai.fi

iQRL

Representation learning



Encoder $z_t = \boxed{f}(e_\theta(o_t))$

Finite Scalar Quantization

Dynamics $\hat{z}_{t+1} = \boxed{f}(z_t + d_\phi(z_t, a_t))$

Latent-state consistency loss

$$\arg\min_{\theta,\phi} \sum_{h=t}^{t+H} \gamma^h \left(\frac{f(\hat{z}_h + d_\phi(\hat{z}_h, a_h))}{\|f(\hat{z}_h + d_\phi(\hat{z}_h, a_h))\|_2} \right)^\top \left(\frac{f(\boxed{e_{\bar{\theta}}}(o_{h+1}))}{\|f(\boxed{e_{\bar{\theta}}}(o_{h+1}))\|_2} \right)$$

Momentum encoder $\bar{\theta} \leftarrow (1-\tau)\bar{\theta} + \tau\theta$

iQRL

Algorithm

i. For i in number of episodes
   i. Collect trajectory $\tau_i = \{o_t, a_t, o_{t+1}, r_t\}_{t=0}^{T}$
   ii. Add trajectory to replay buffer $\mathscr{D} \leftarrow \mathscr{D} \cup \tau_i$
   iii. For $T \times r_{\text{utd}}$ steps
      i. Sample batch from replay buffer $\mathscr{D}$
      ii. One encoder update
      iii. One critic update
      iv. One actor update

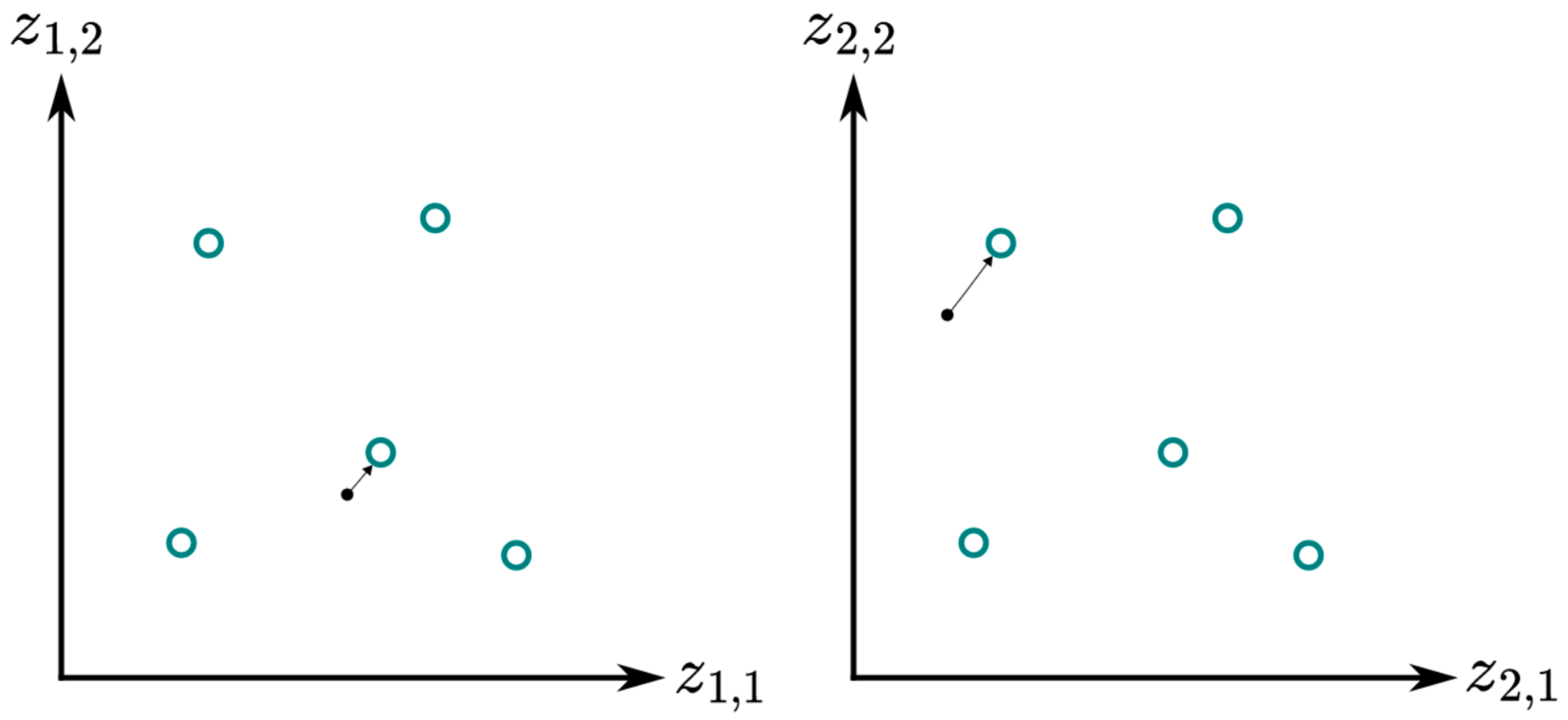
Finite Scalar Quantization

FSQ does not learn a codebook

It's pre-specified by hyperparameters

Vector Quantization

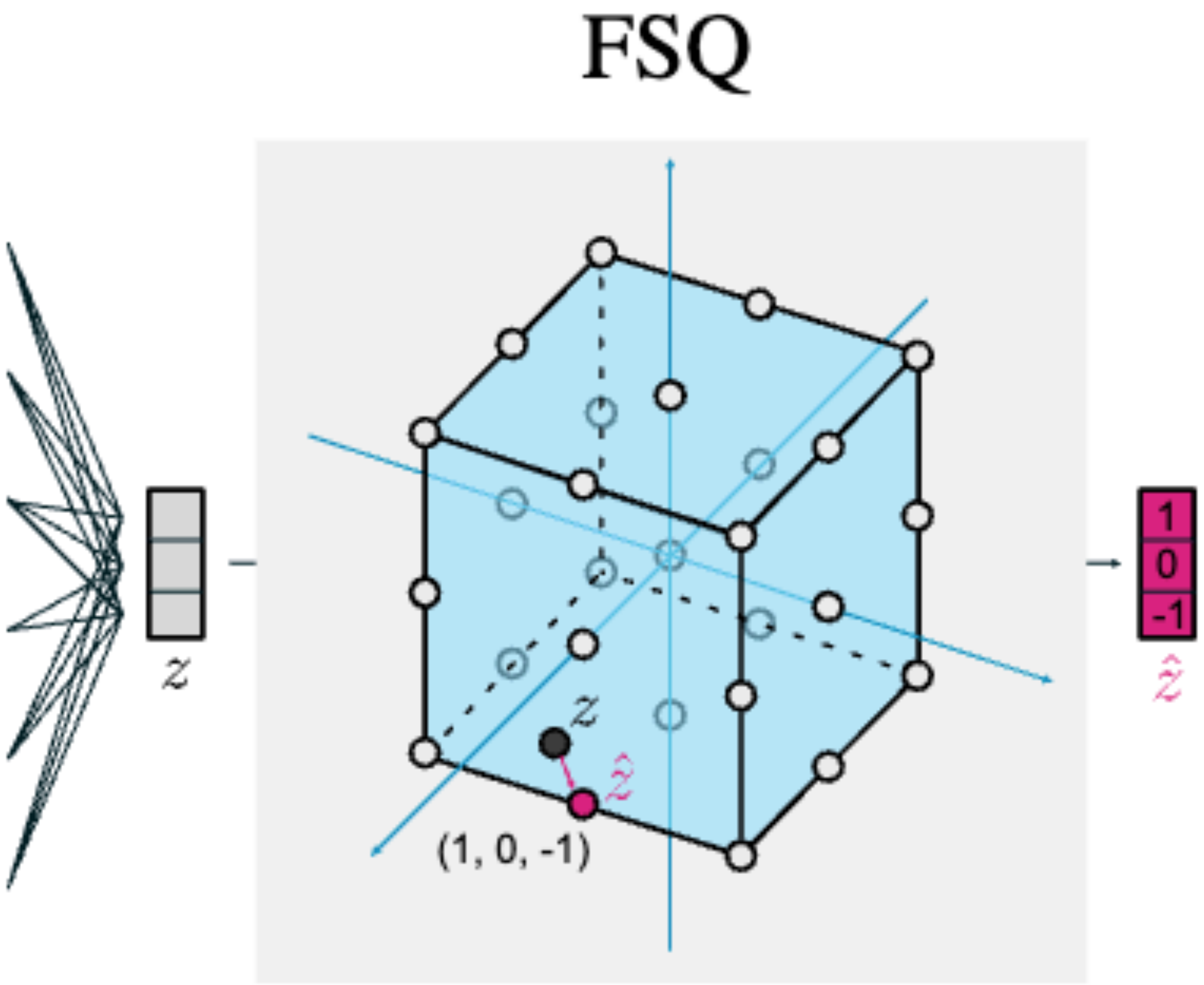
Finite Scalar Quantization



FCAI

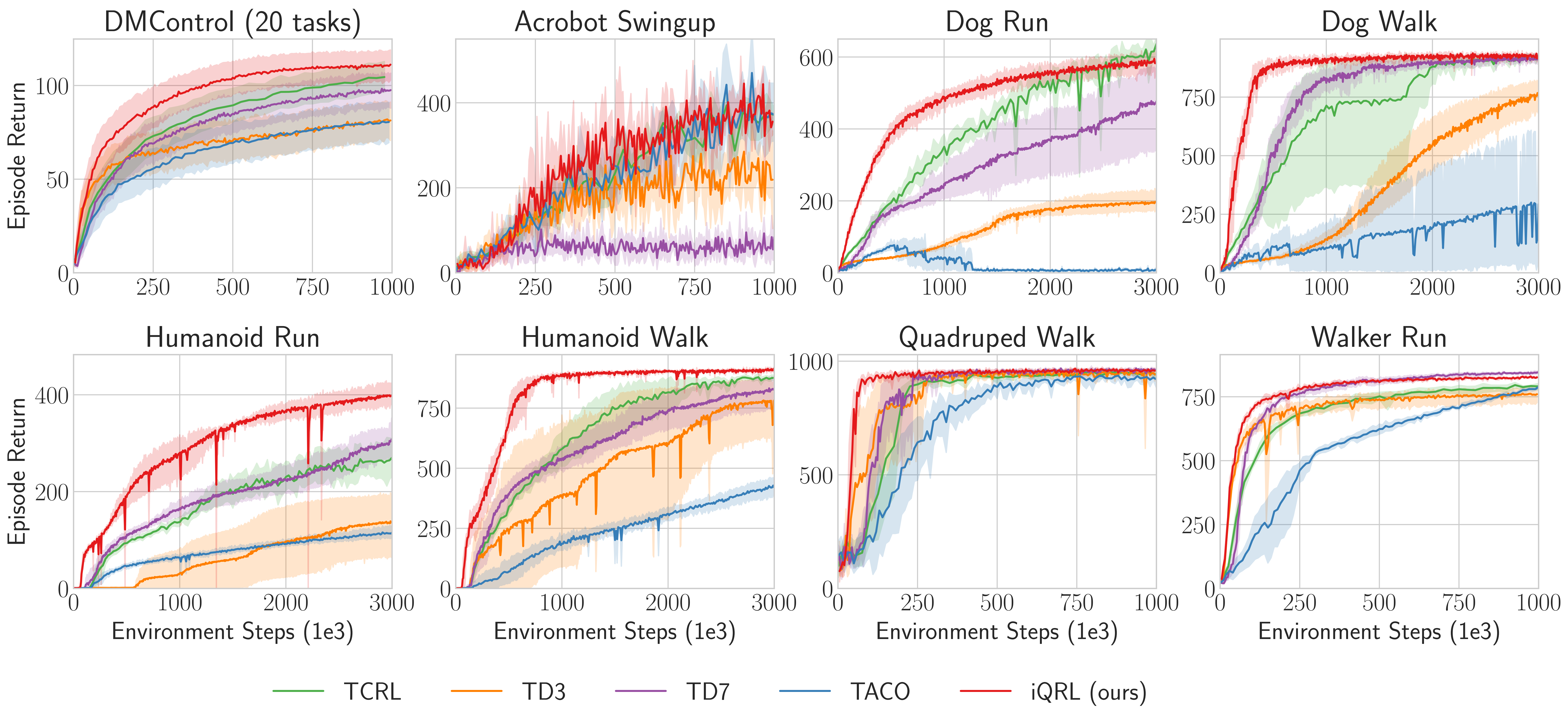
fcai.fi

Finite Scalar Quantization



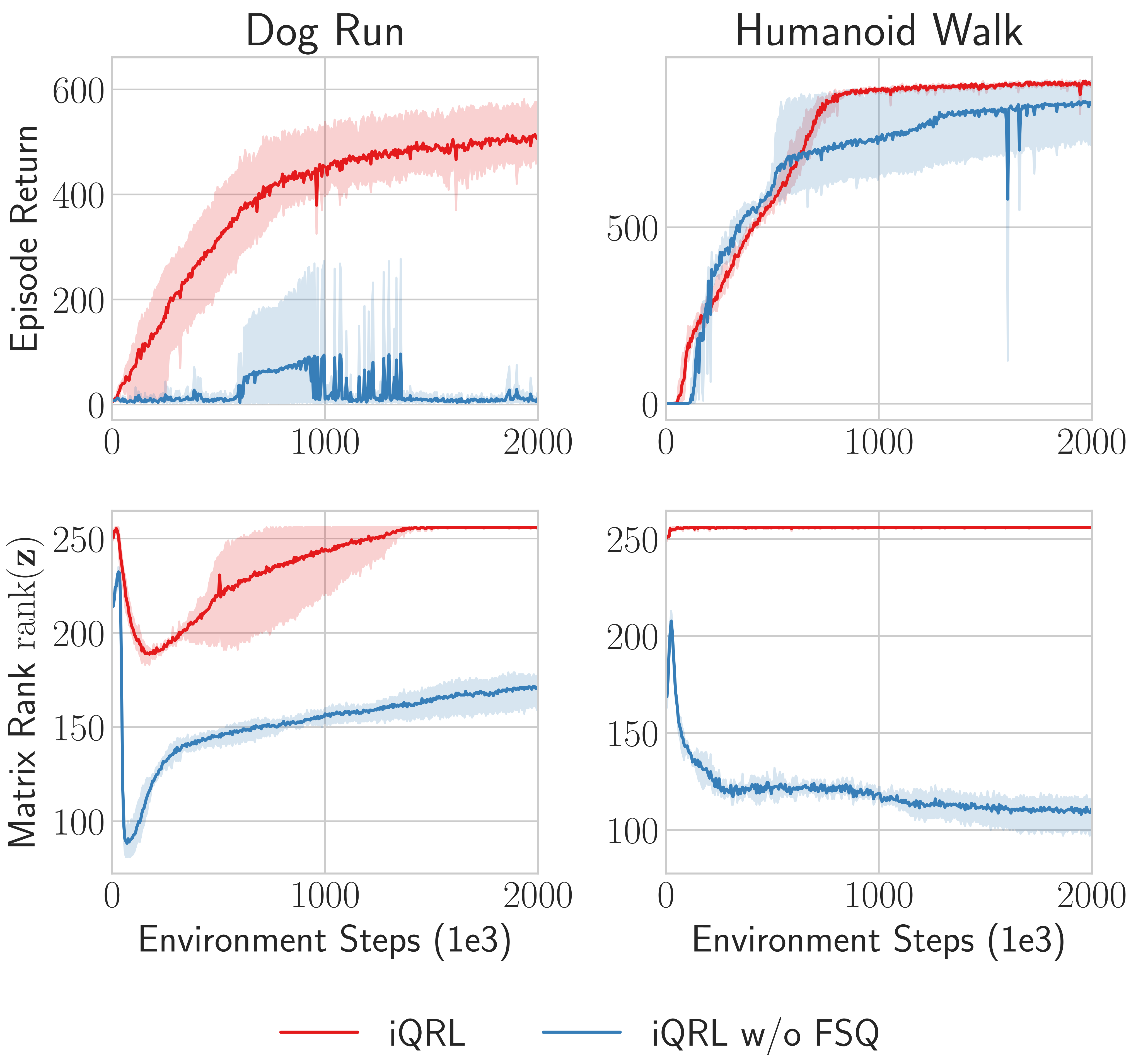
Results

Strong Performance in DMControl



Results

FSQ (Empirically) Prevents Dimensional Collapse

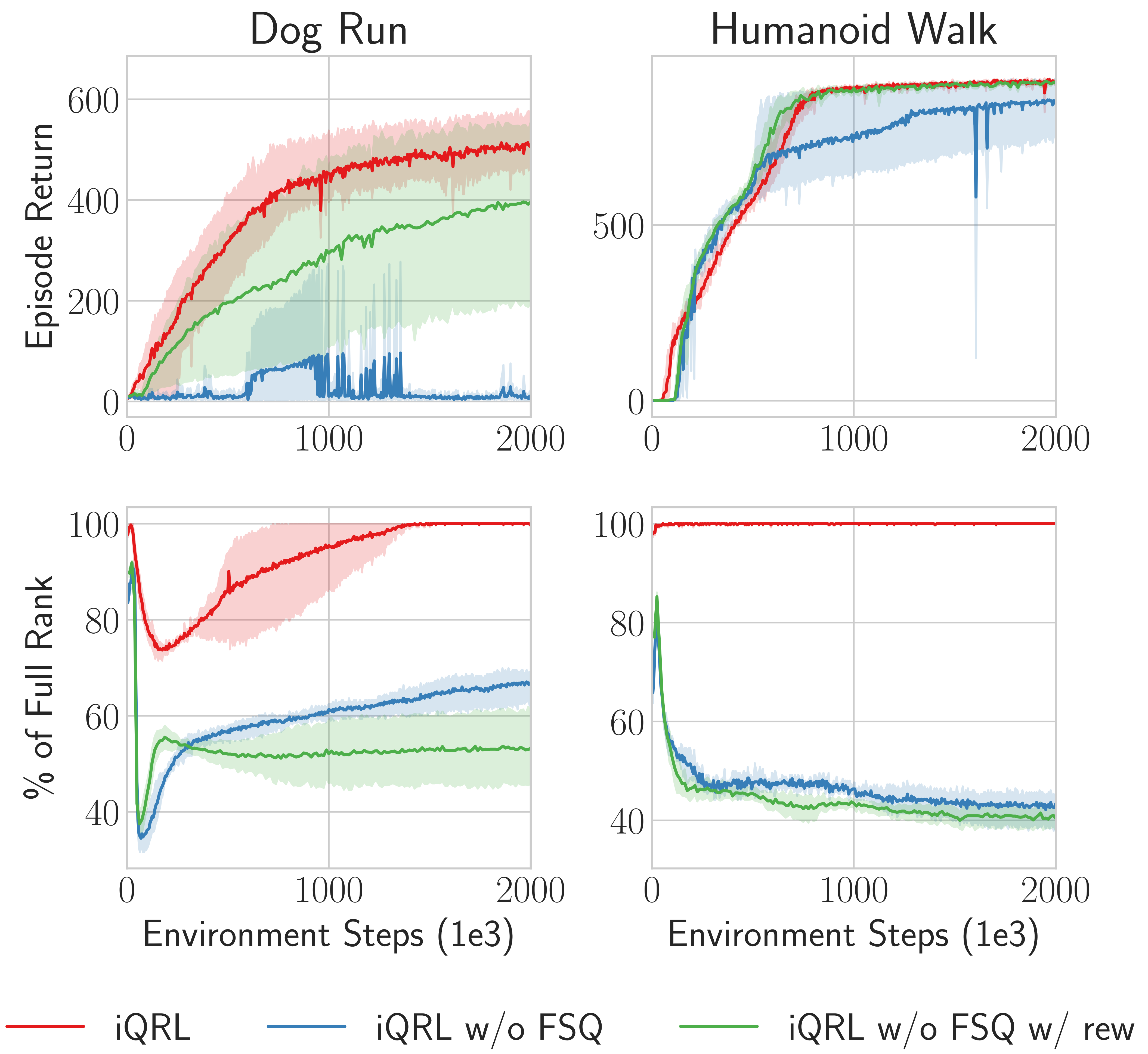


FCAI

fcai.fi

Results

Reward Prediction Helps a Little

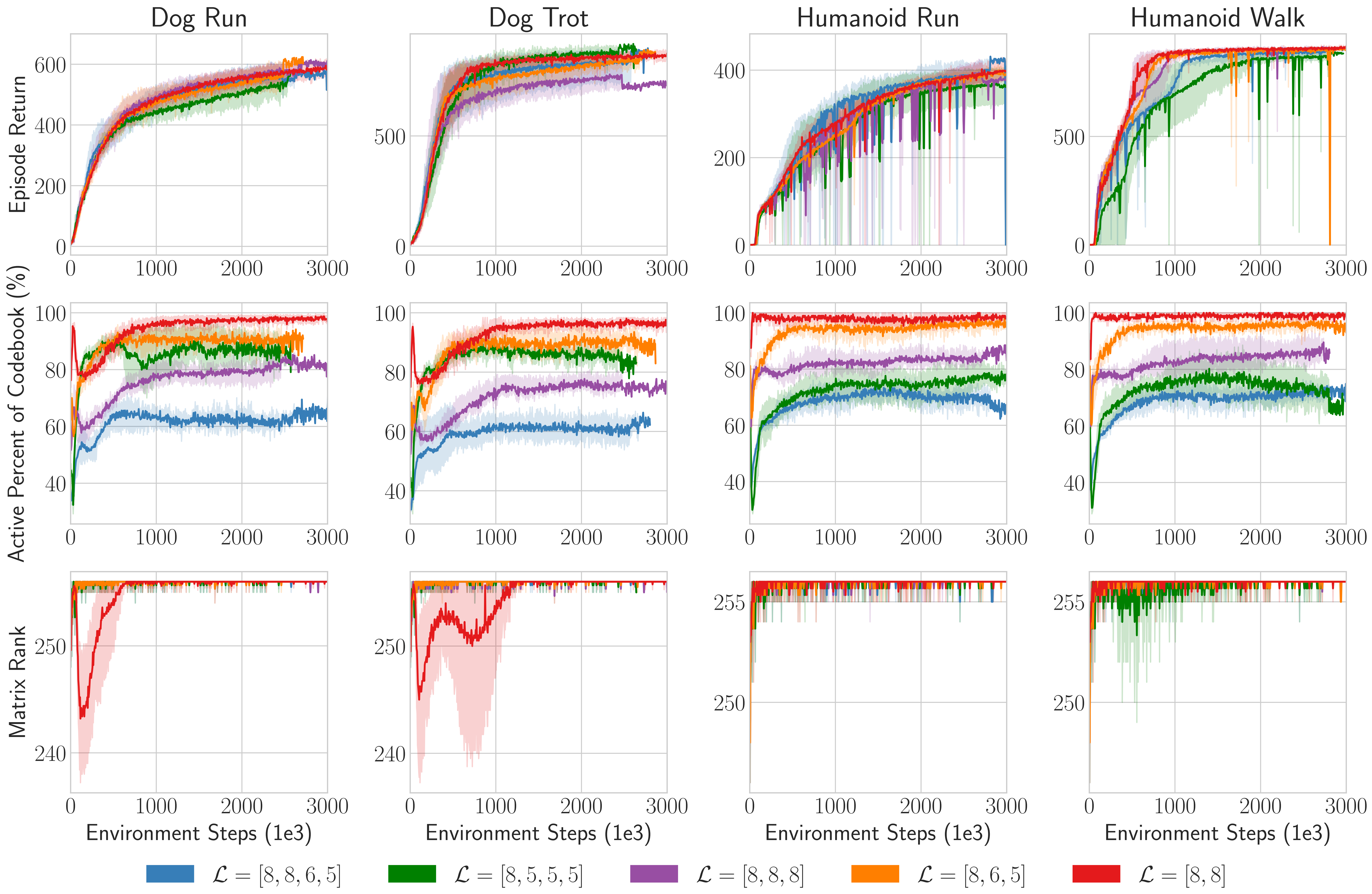


FCAI

fcai.fi

Hyperparameter Analysis

Robust to Codebook Size

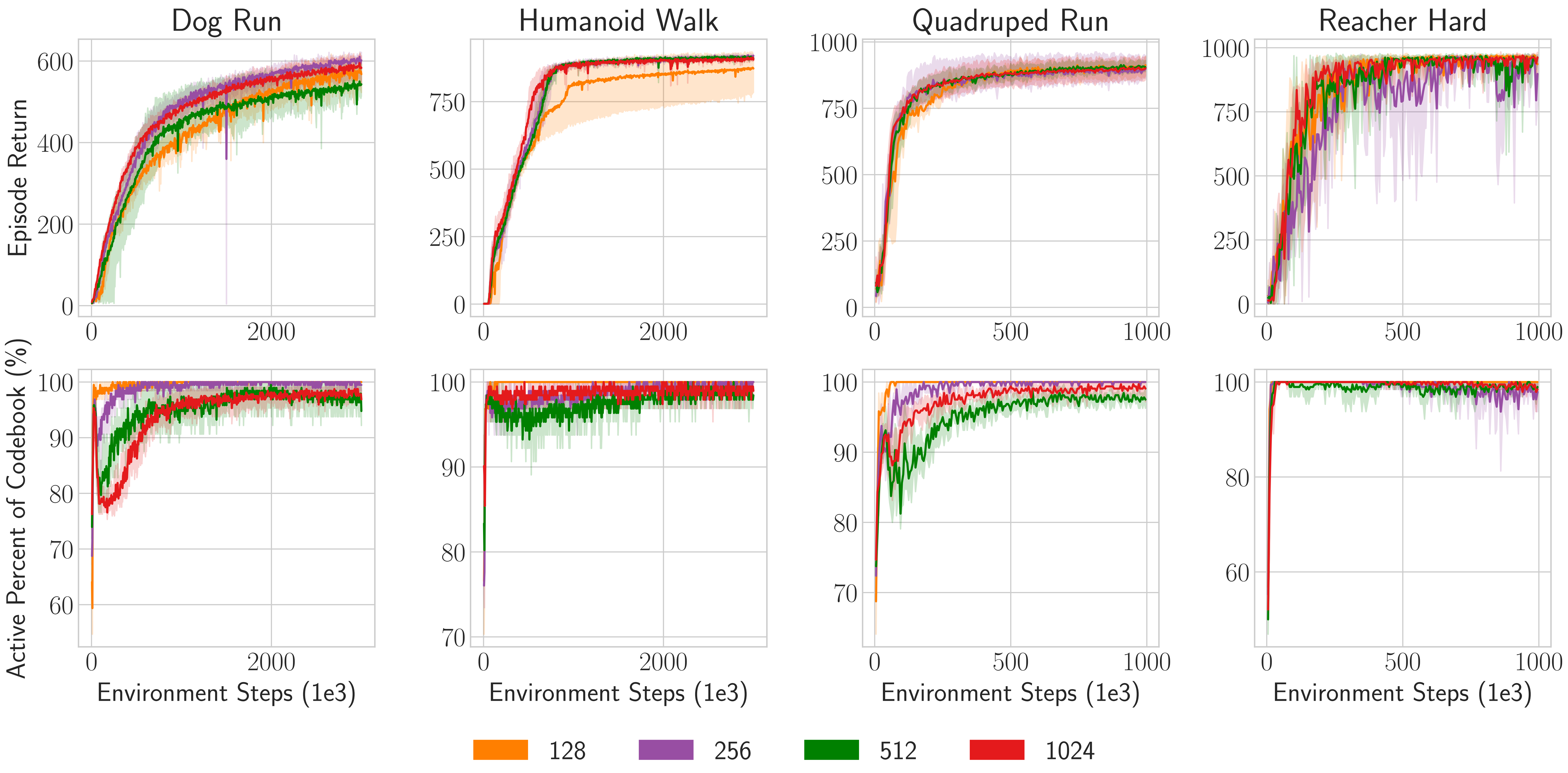


FCAI

fcai.fi

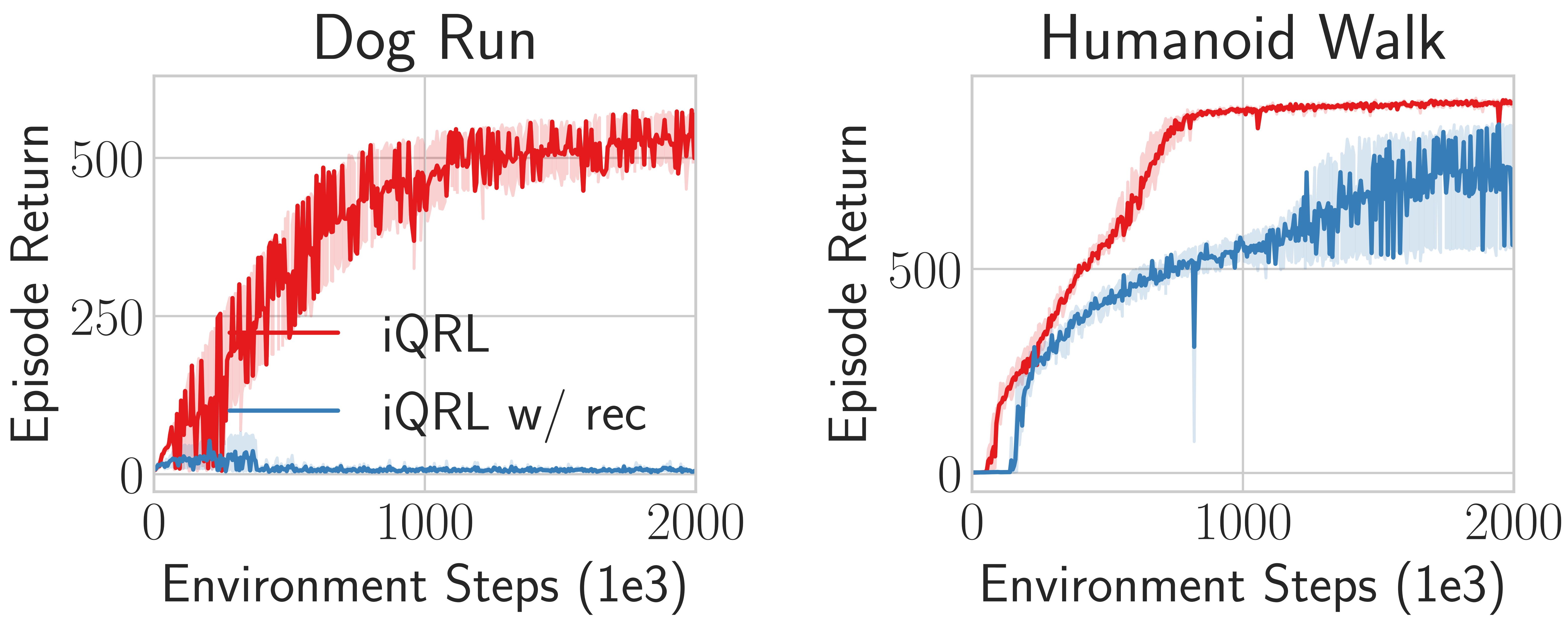
Hyperparameter Analysis

Robust to Size of Latent State



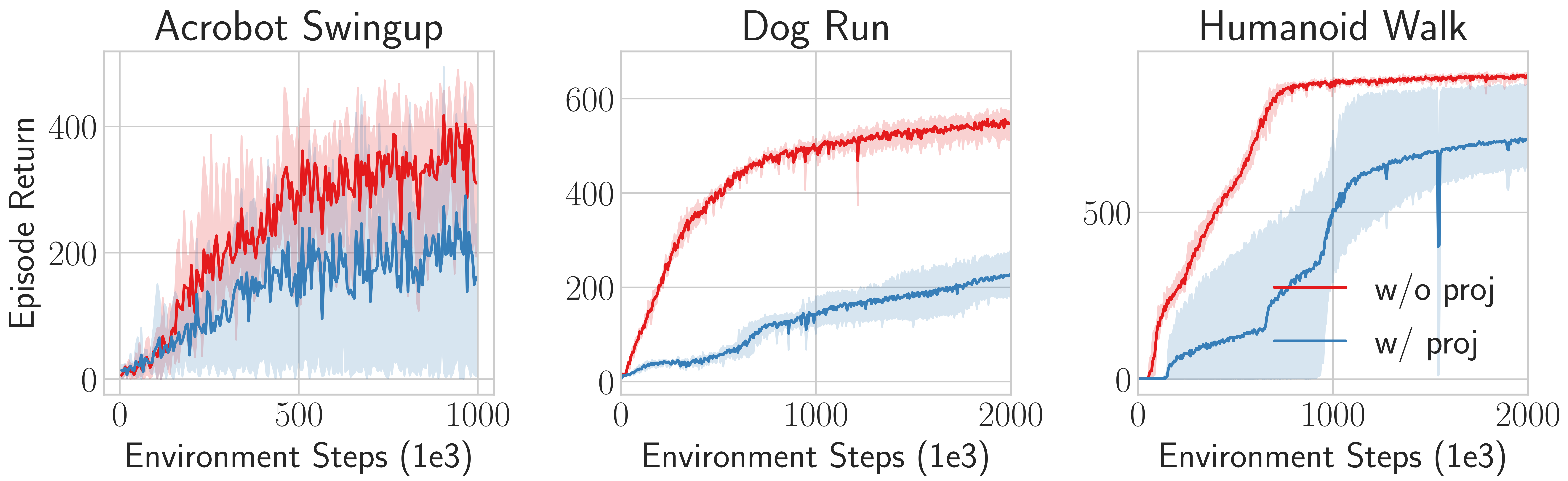
Further Insights

Reconstruction Loss Harms Performance



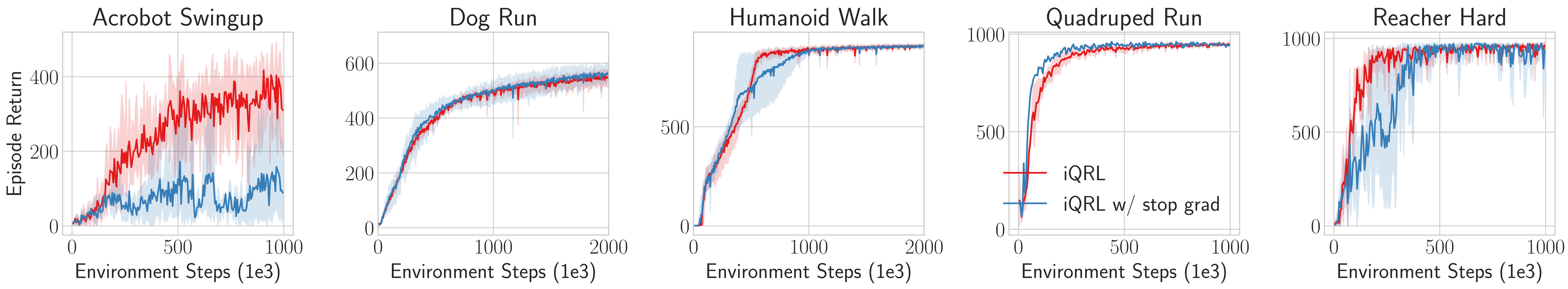
Further Insights

Projection Head Harms Sample Efficiency



Further Insights

Momentum Encoder > Stop Gradient

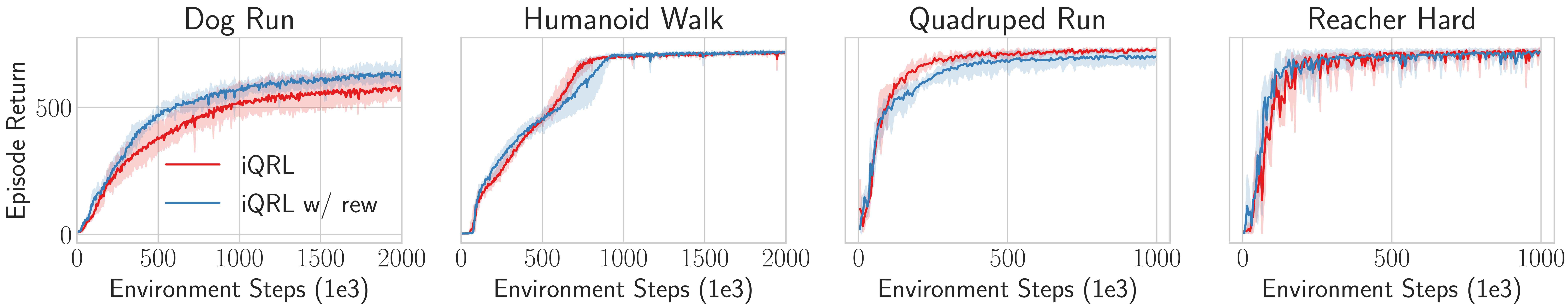


FCAI

fcai.fi

Further Insights

Reward Head (Only) Improves a Little



Insights and Takeaways

iQRL

- Straightforward
- Compatible with any model-free RL algorithm
- Fast (no decision-time planning)
- Strong performance in DMControl
- Representation is task agnostic
- Quantization (empirically) prevents dimensional collapse

Insights

- Learning a high-dimensional latent state (d=512/1024) makes Q-learning easier…
- Difficulty of Q-learning is due to complex dynamics, not high-dimensional observations

Email: aidan.scannell@aalto.fi

Website: www.aidanscannell.com