Enhancing Interpretability in Deep Reinforcement Learning through Semantic Clustering

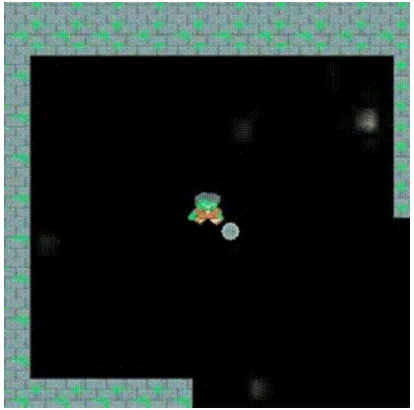
Liang Zhang, Justin Lieffers, Adarsh Pyarelal
College of Information Science, University of Arizona

What is semantic clustering in DRL?

- State sequences with similar semantics naturally group into coherent clusters that reflect behaviors and persist across time.

Procgen domains

making small jumps



touching the mushroom

Limitations of prior work

- Post-hoc t-SNE is unstable under state count and random seeds, which is hard to reproduce.
- Fixed-scene Atari emphasizes pixel similarity rather than semantics.
- No automatic clustering: heavy manual grouping and feature engineering.

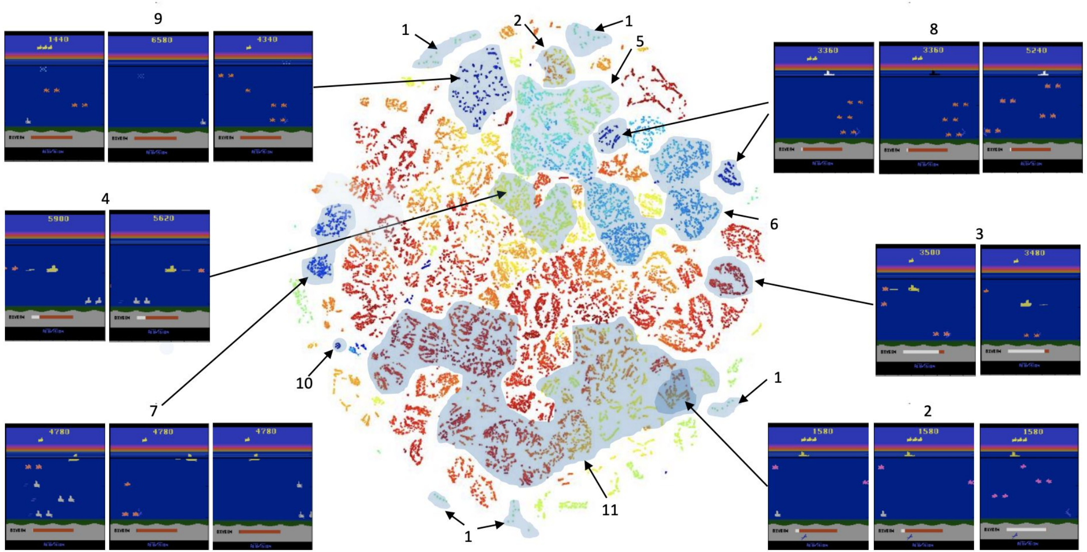


Figure 3. Seaquest aggregated states on the t-SNE map, colored by value function estimate.

From Zahavy et al., Graying the black box: Understanding DQNs, ICML 2016.

Contributions

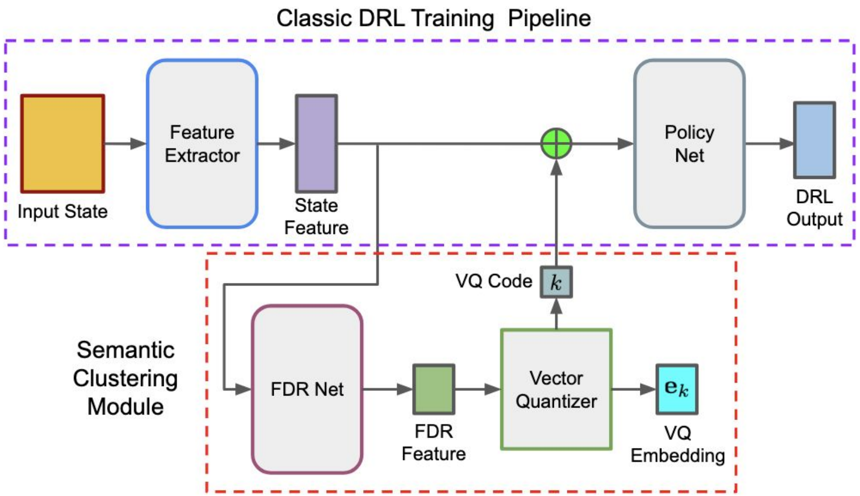
1. Systematic study of DRL semantic clustering in Procgen (beyond fixed-scene Atari).
2. An online semantic clustering module for real-time feature dimensionality reduction and online clustering.
3. New analysis methods that reveal internal semantic structure, hierarchical policy organization, and potential risks.

Procgen domains

Architecture overview

- Pipeline: state → feature extractor $f(s)$ → FDR 2-D mapping $g(\cdot)$ → vector quantizer with codebook e_k → discrete code k → fuse k with features → conditional policy head.
- Control factor λ_{ctrl} limits early interference and ramps up clustering influence as learning progresses.

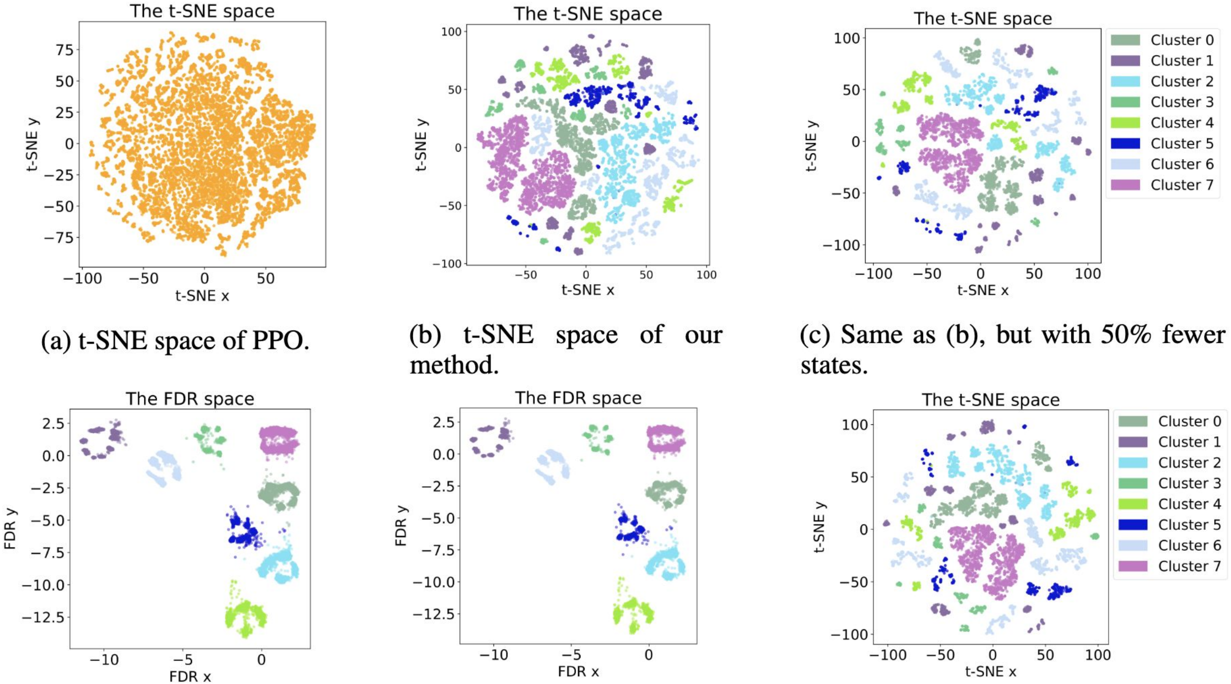
Overview of our architecture



Total objective: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{DRL}} + \lambda_{\text{ctrl}} \left(w_{\text{FDR}} \, \mathcal{L}_{\text{FDR}} + w_{\text{VQ-VAE}} \, \mathcal{L}'_{\text{VQ-VAE}} \right).$

Results I: stability, semantics, consistency

- Stability vs t-SNE: FDR forms well-separated, stable clusters under halved samples and different seeds; t-SNE is sensitive and fragmentary.



(a) t-SNE space of PPO.

(b) t-SNE space of our method.

(c) Same as (b), but with 50% fewer states.

(d) FDR space of our method.

(e) Same as (d), but with 50% fewer states.

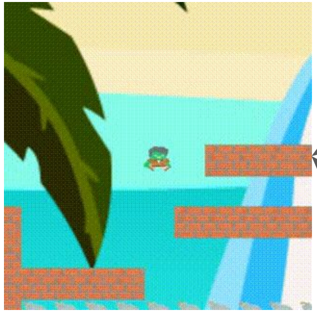
(f) Same as (c), but with a different random seed.

Visualization of features in t-SNE and FDR spaces using PPO and our method.

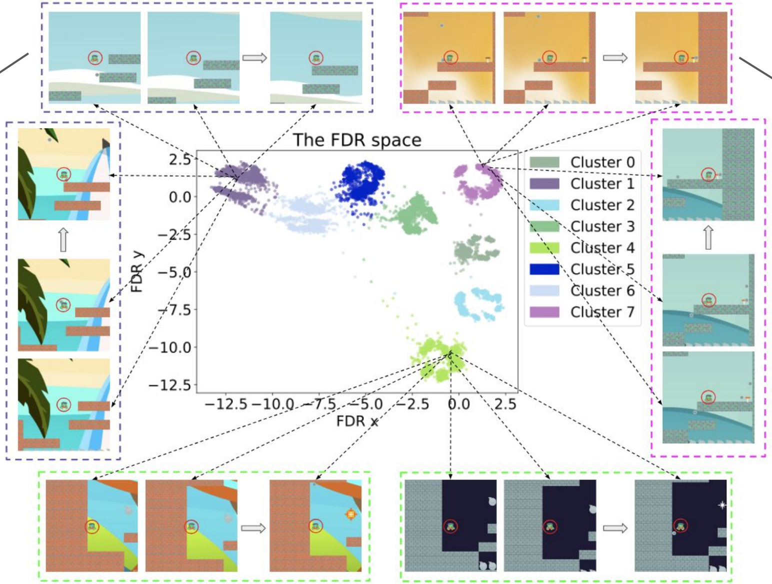
Results I: stability, semantics, consistency

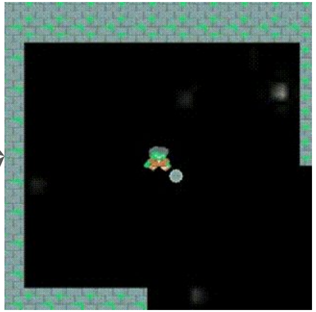
- Cluster semantics: within-cluster state strips show temporal coherence. For dynamic videos and per-cluster behavior descriptions, please see our code repository.

State examples in the proposed FDR space



Cluster 1: making small jumps

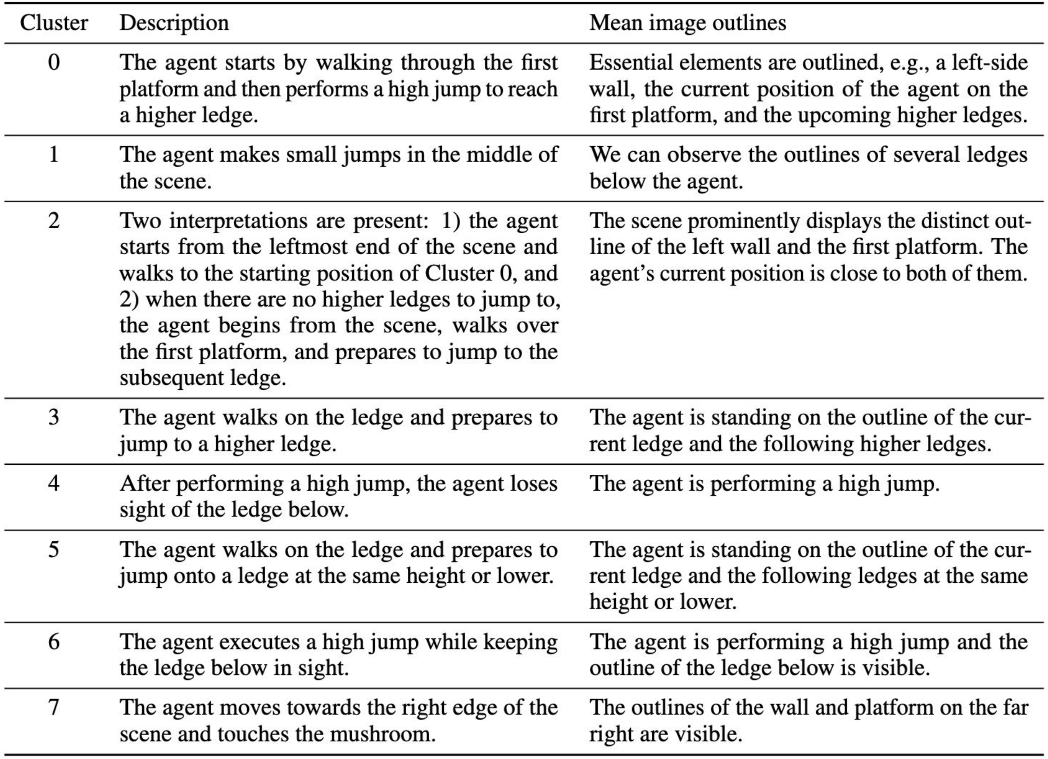




Cluster 7: touching the mushroom

Results I: stability, semantics, consistency

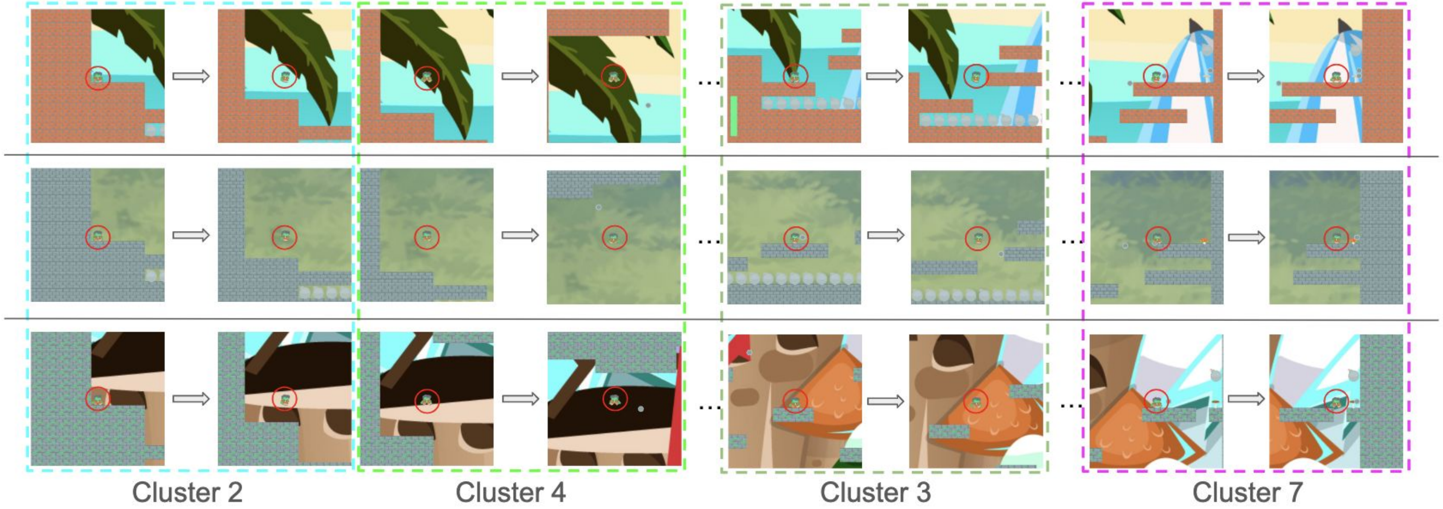
Cluster descriptions and mean image outlines for the Ninja game



Cluster	Description	Mean image outlines
0	The agent starts by walking through the first platform and then performs a high jump to reach a higher ledge.	Essential elements are outlined, e.g., a left-side wall, the current position of the agent on the first platform, and the upcoming higher ledges.
1	The agent makes small jumps in the middle of the scene.	We can observe the outlines of several ledges below the agent.
2	Two interpretations are present: 1) the agent starts from the leftmost end of the scene and walks to the starting position of Cluster 0, and 2) when there are no higher ledges to jump to, the agent begins from the scene, walks over the first platform, and prepares to jump to the subsequent ledge.	The scene prominently displays the distinct outline of the left wall and the first platform. The agent's current position is close to both of them.
3	The agent walks on the ledge and prepares to jump to a higher ledge.	The agent is standing on the outline of the current ledge and the following higher ledges.
4	After performing a high jump, the agent loses sight of the ledge below.	The agent is performing a high jump.
5	The agent walks on the ledge and prepares to jump onto a ledge at the same height or lower.	The agent is standing on the outline of the current ledge and the following ledges at the same height or lower.
6	The agent executes a high jump while keeping the ledge below in sight.	The agent is performing a high jump and the outline of the ledge below is visible.
7	The agent moves towards the right edge of the scene and touches the mushroom.	The outlines of the wall and platform on the far right are visible.

Results I: stability, semantics, consistency

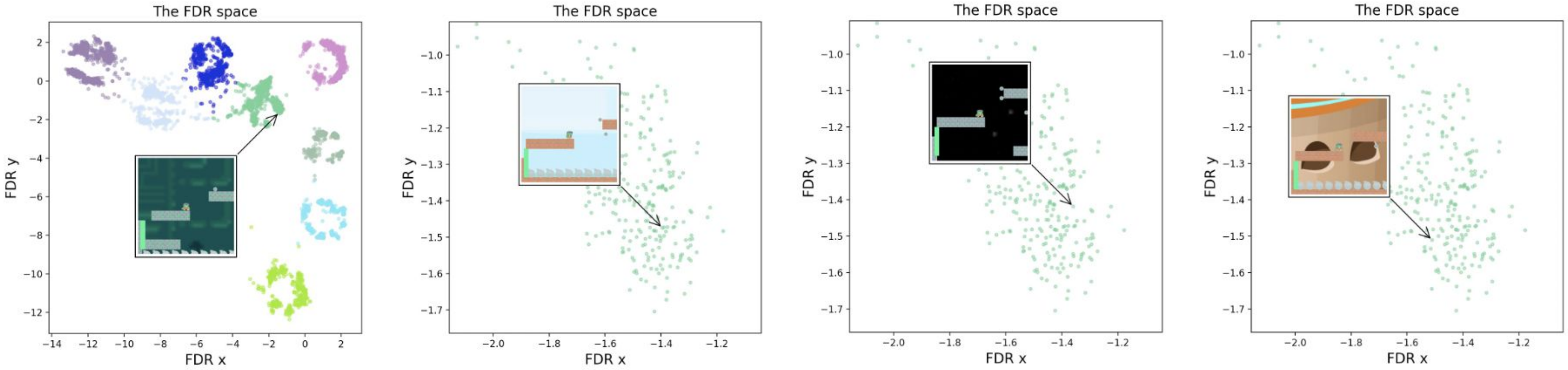
- Cross-episode consistency: the same clusters align across episodes (e.g., Ninja), indicating reusable skills.



Three episodes from the Ninja game.

Results II: analysis, human evaluation, performance

- Policy forensics: interactive FDR viewer reveals sub-clusters


The FDR space

The FDR space

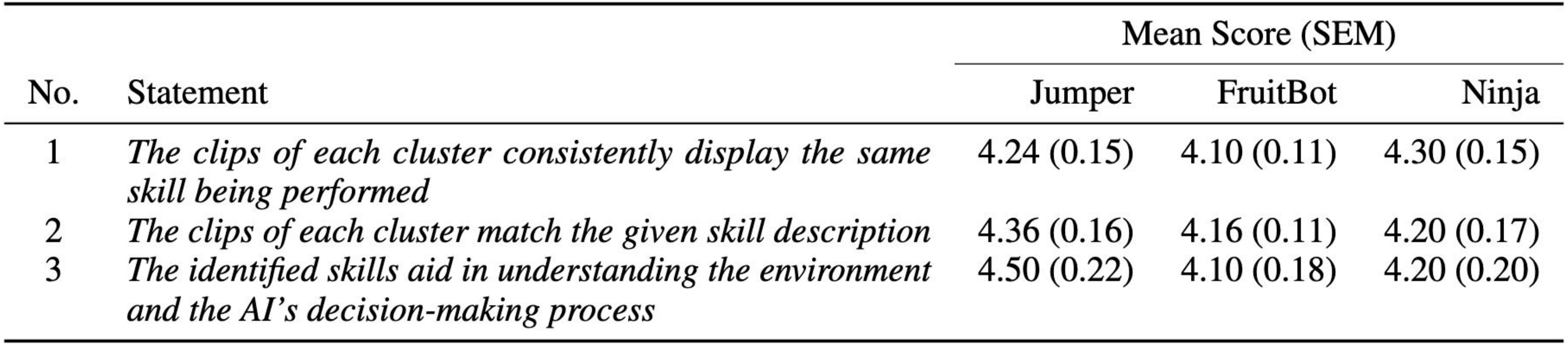
The FDR space

The FDR space

Hover examples in the FDR space of Ninja.

Results II: analysis, human evaluation, performance

- Human evaluation (Table 2): Likert means ≈4.1–4.5/5 for “same skill,” “matches description,” and “helps understand decisions.”

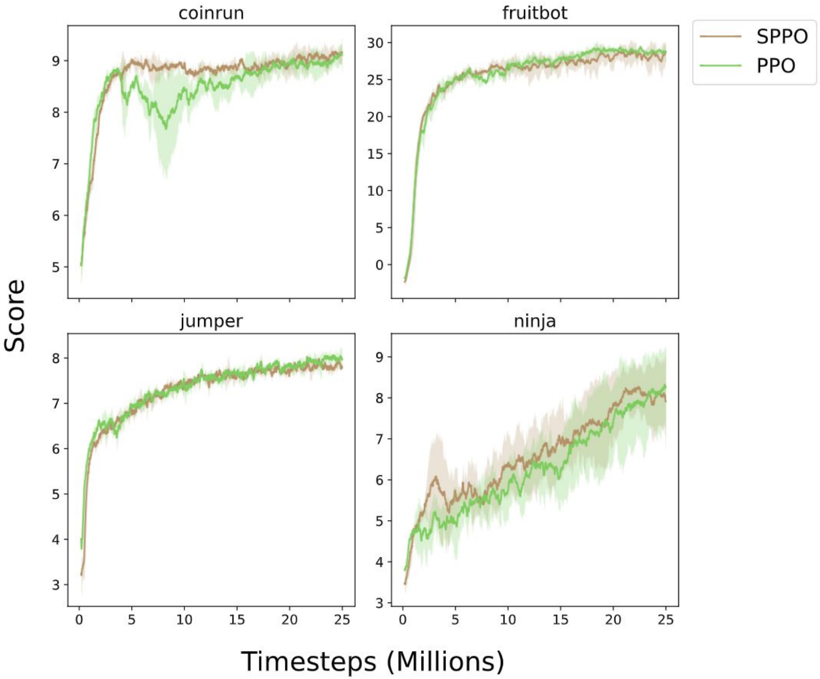


		Mean Score (SEM)		
No.	Statement	Jumper	FruitBot	Ninja
1	*The clips of each cluster consistently display the same skill being performed*	4.24 (0.15)	4.10 (0.11)	4.30 (0.15)
2	*The clips of each cluster match the given skill description*	4.36 (0.16)	4.16 (0.11)	4.20 (0.17)
3	*The identified skills aid in understanding the environment and the AI’s decision-making process*	4.50 (0.22)	4.10 (0.18)	4.20 (0.20)

Human evaluation results

Results II: analysis, human evaluation, performance

- Minimal performance impact: comparable PPO scores; varying the number of embeddings does not degrade performance.



Performance curves

Conclusion

- An online FDR+VQ module learns a stable, human-aligned semantic space and discrete codes during training, enabling cross-episode checks, semantic segmentation, and policy forensics with minimal performance cost.

Paper Link:

Code Link: