Sliding Puzzles Gym: A Scalable Benchmark for State Representation in Visual RL

Bryan L. M. de Oliveira[1,2], Luana G. B. Martins[1], Bruno Brandão[1,2], Murilo L. da Luz[1,2], Telma W. de L. Soares[1,2], Luckeciano C. Melo[1,3]

ICML 2025

[1]Advanced Knowledge Center for Immersive Technologies (AKCIT)
[2]Institute of Informatics, Federal University of Goiás
[3]OATML, University of Oxford

Correspondence to: bryanlincoln@discente.ufg.br
Code: https://github.com/bryanoliveira/sliding-puzzles-gym

Targeted Visual RL Evaluation

- The Challenge: How do we measure an RL agent's ability to **see and understand** visual content, separate from other skills?

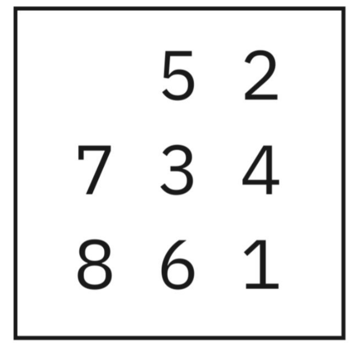
Targeted Visual RL Evaluation

- The Challenge: How do we measure an RL agent's ability to **see and understand** visual content, separate from other skills?
  - Existing benchmarks (e.g., Atari, ProcGen, DM Control) are great, but **they mix different challenges together:** representation learning, policy learning, dynamics learning.

Targeted Visual RL Evaluation

- The Challenge: How do we measure an RL agent's ability to **see and understand** visual content, separate from other skills?
    - Existing benchmarks (e.g., Atari, ProcGen, DM Control) are great, but **they mix different challenges together:** representation learning, policy learning, dynamics learning.
- The Gap: There's no systematic way to **isolate and scale only the visual representation challenge**.

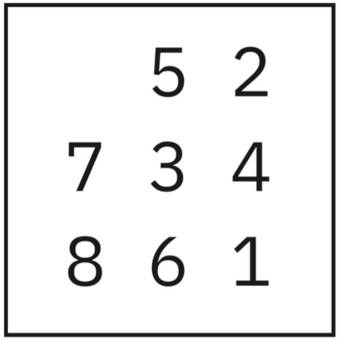
The Sliding Puzzles Gym (SPGym)



State

The Sliding Puzzles Gym (SPGym)

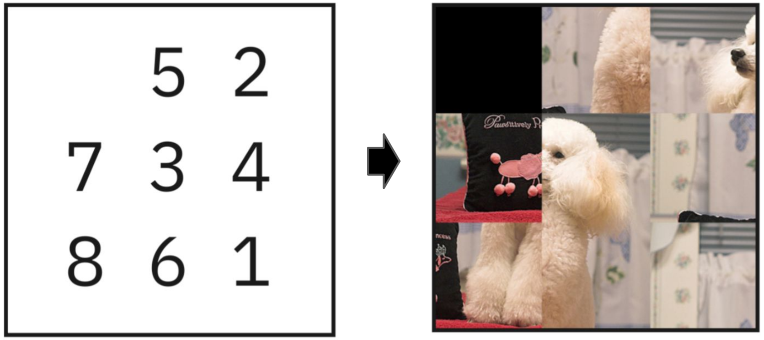
- **Our Solution:** Isolate the visual challenge using the classic 8-puzzle.



State

The Sliding Puzzles Gym (SPGym)

- **Our Solution:** Isolate the visual challenge using the classic 8-puzzle.
    - Tiles are **patches from an image**.

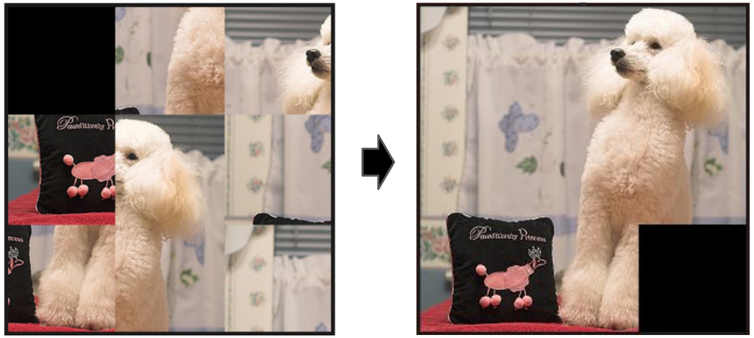


State

Image Overlay

The Sliding Puzzles Gym (SPGym)

- **Our Solution:** Isolate the visual challenge using the classic 8-puzzle.
  - Tiles are **patches from an image**.
  - The task is **always the same**.

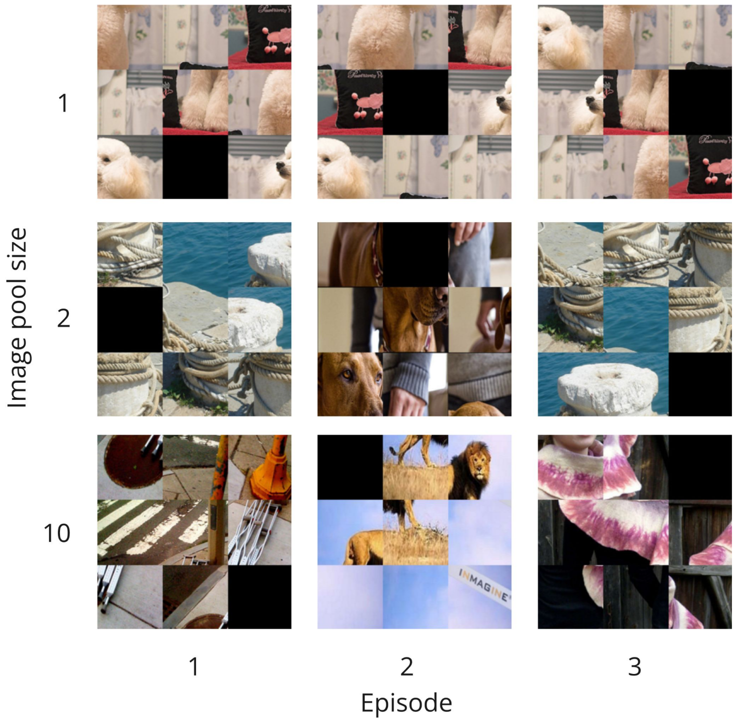


Initial State

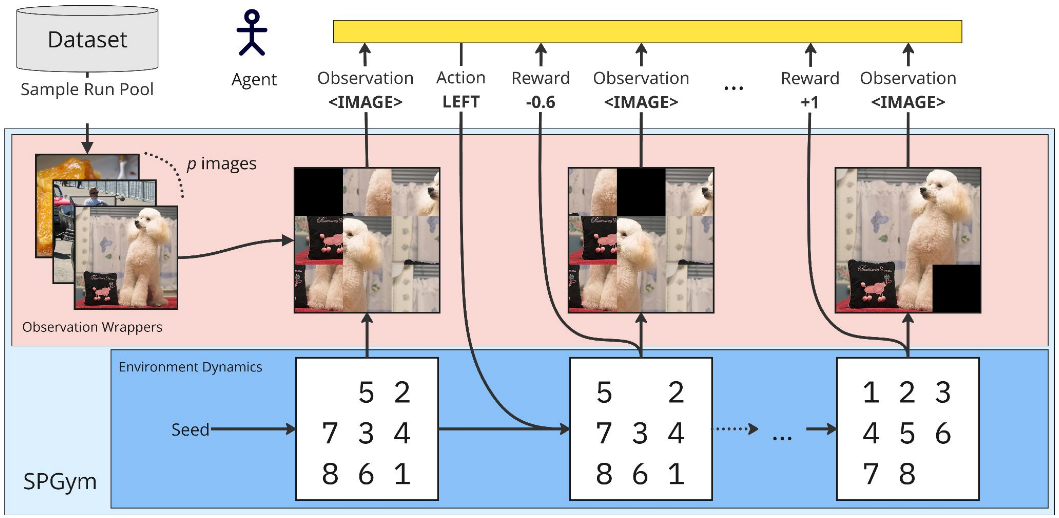
Goal State

The Sliding Puzzles Gym (SPGym)

- **Our Solution:** Isolate the visual challenge using the classic 8-puzzle.
  - Tiles are **patches from an image**.
  - The task is **always the same**.
  - Visual diversity is controlled by increasing the **pool of images**.

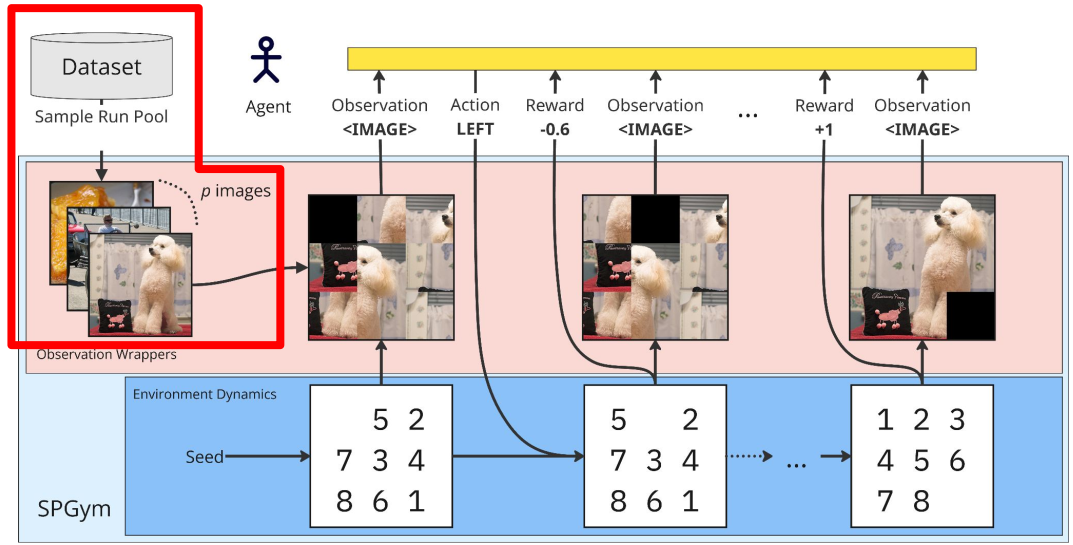


Method Overview



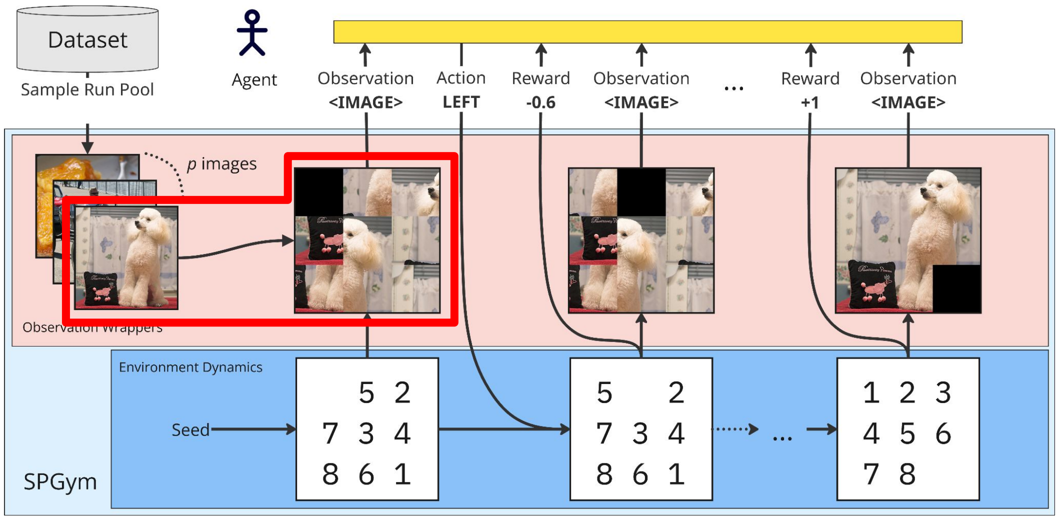
Method Overview

- At run start: Sample images from the **dataset** to form an **image pool.**



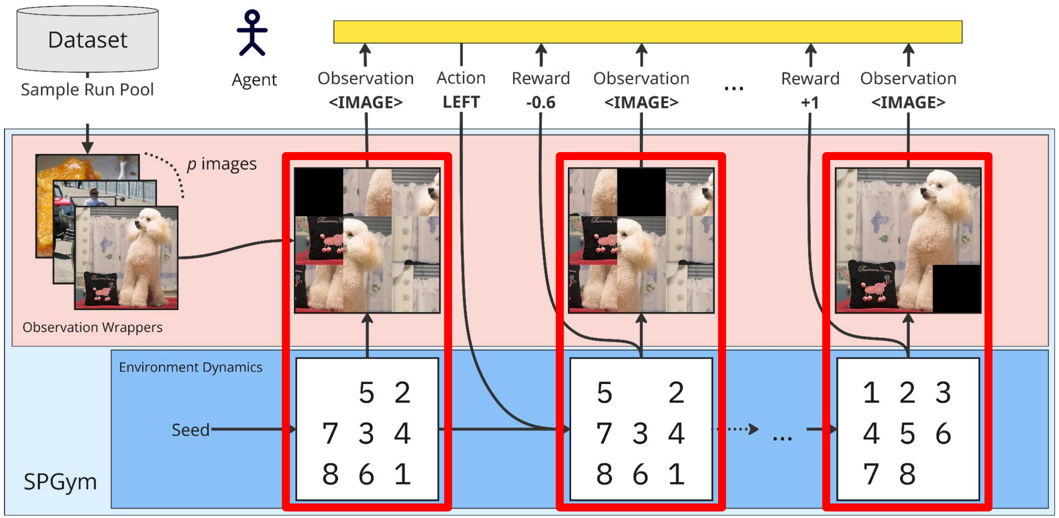
Method Overview

- At run start: Sample images from the **dataset** to form an **image pool**.
- At episode start: Sample an image from the pool and split it into **indexed patches**.



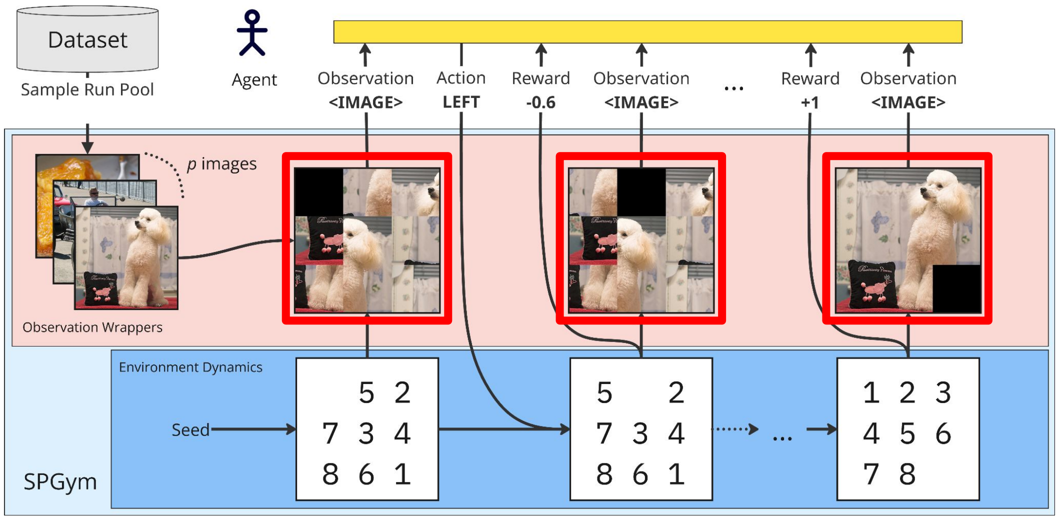
Method Overview

- At run start: Sample images from the **dataset** to form an **image pool**.
- At episode start: Sample an image from the pool and split it into **indexed patches**.
- **Overlay** patches onto the puzzle state.



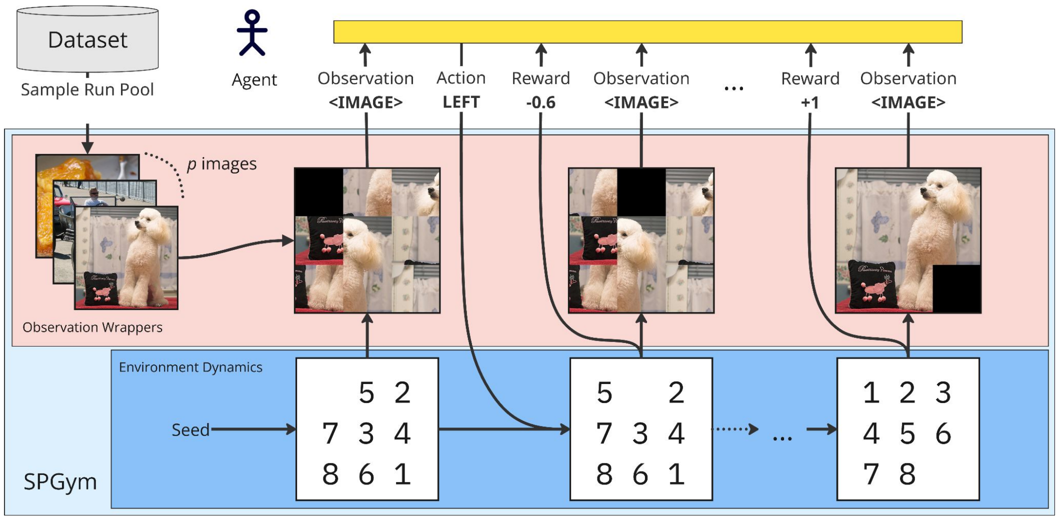
Method Overview

- At run start: Sample images from the **dataset** to form an **image pool**.
- At episode start: Sample an image from the pool and split it into **indexed patches**.
- **Overlay** patches onto the puzzle state.



Method Overview

- At run start: Sample images from the **dataset** to form an **image pool**.
- At episode start: Sample an image from the pool and split it into **indexed patches**.
- **Overlay** patches onto the puzzle state.
- Visual complexity controls: Image **pool** and **grid sizes**.



Experimental Setup

- Goal: measure how modern RL agents **handle increasing visual diversity**.

Experimental Setup

- Goal: measure how modern RL agents **handle increasing visual diversity**.
- Environment: **3x3 grids with images** from ImageNet.

Experimental Setup

- Goal: measure how modern RL agents **handle increasing visual diversity**.
- Environment: **3x3 grids with images** from ImageNet.
- Independent Variable: Images in the training pool (from 1 up to 100).

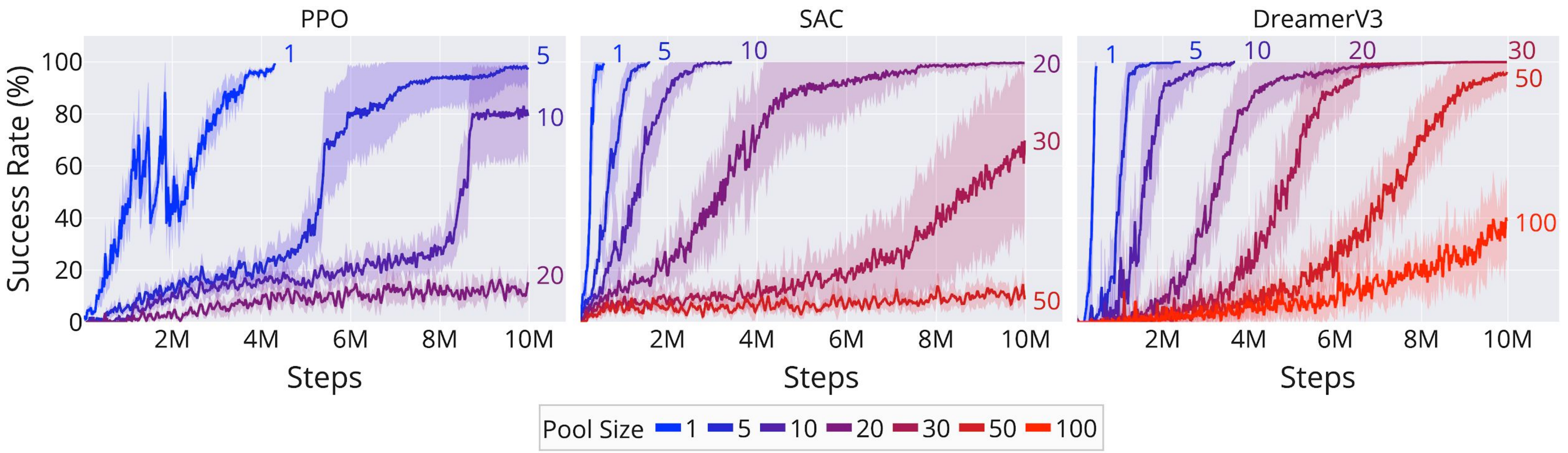
Experimental Setup

- Goal: measure how modern RL agents **handle increasing visual diversity**.
- Environment: **3x3 grids with images** from ImageNet.
- Independent Variable: Images in the training pool (from 1 up to 100).
- Algorithms: PPO, SAC and DreamerV3 with multiple variants.

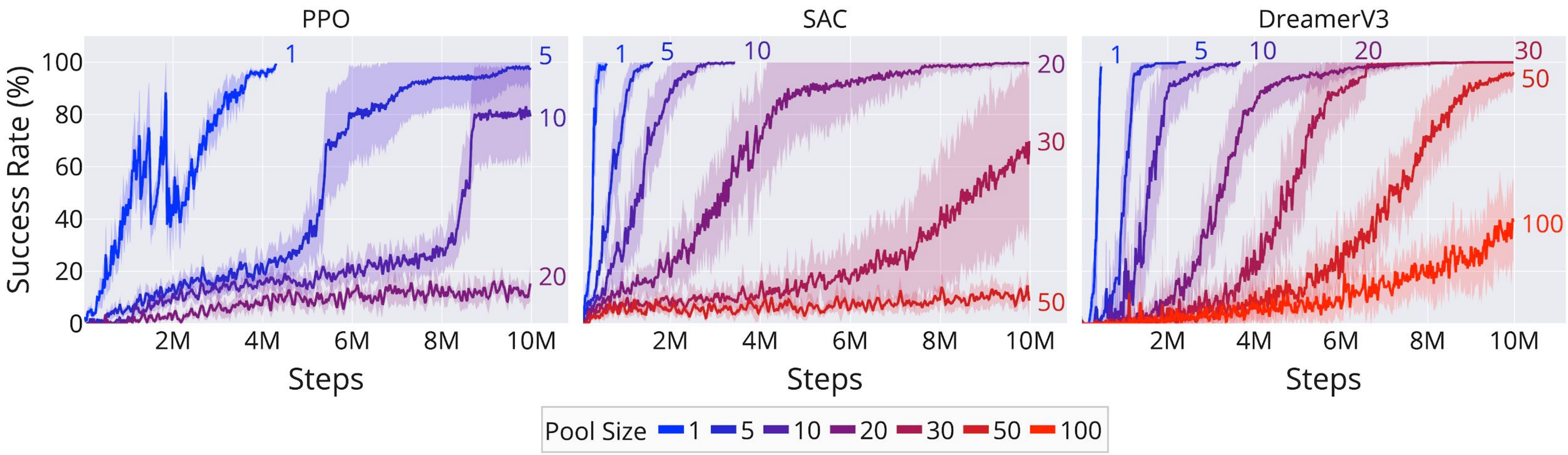
Experimental Setup

- Goal: measure how modern RL agents **handle increasing visual diversity**.
- Environment: **3x3 grids with images** from ImageNet.
- Independent Variable: Images in the training pool (from 1 up to 100).
- Algorithms: PPO, SAC and DreamerV3 with multiple variants.
- Primary Metric: **Sample Efficiency** (steps to 80% success rate).

Results: Performance & Scaling

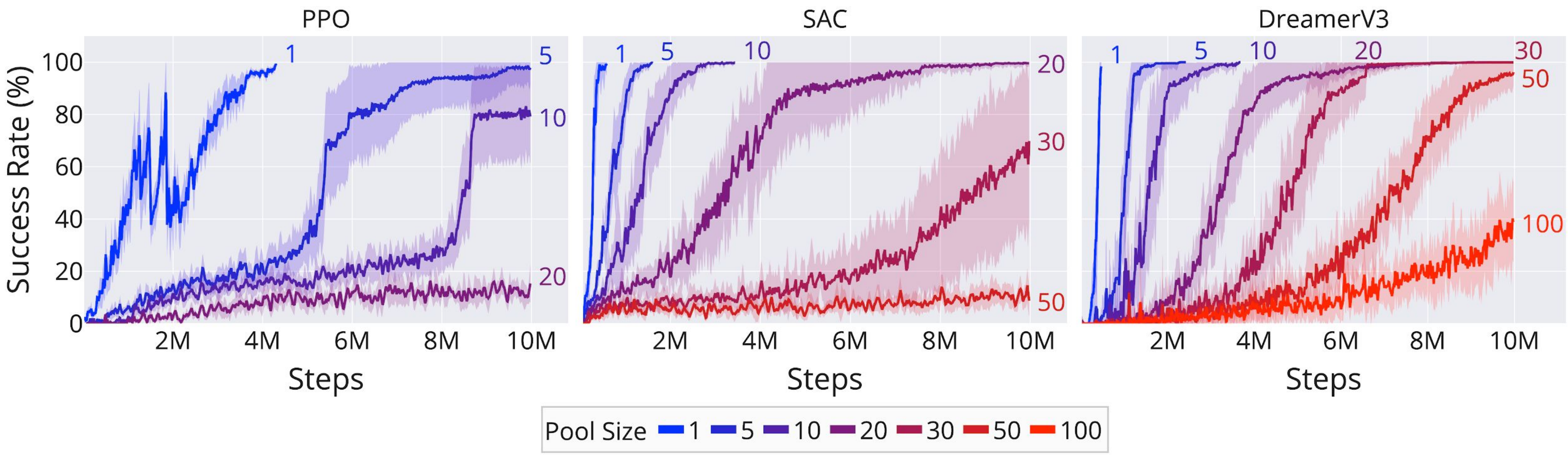


Results: Performance & Scaling



- All agents take **longer** to learn as the image pool **grows**.

Results: Performance & Scaling



- All agents take **longer** to learn as the image pool **grows**.
- DreamerV3 is the **most robust**, likely due to its world model.

Results: Performance of Variants

Table 1. **Million steps to reach 80% success rate across pool sizes.** Lower is better. Best performing variant for each algorithm and pool size is highlighted in bold.

Agent	Pool 1	Pool 5	Pool 10
PPO	1.75±0.44	7.80±1.08	9.73±0.36
PPO + PT (ID)	**0.95±0.21**	**5.55±1.22**	**9.17±1.10**
PPO + PT (OOD)	1.34±0.42	7.03±1.07	9.70±0.41
SAC	0.33±0.07	0.91±0.12	2.03±0.38
SAC + RAD	**0.24±0.03**	**0.42±0.06**	**0.82±0.18**
SAC + CURL	0.46±0.10	1.56±0.31	5.24±1.92
SAC + SPR	2.09±0.81	3.68±1.68	10.00±0.00
SAC + DBC	0.99±0.25	1.12±0.22	2.13±0.41
SAC + AE	1.04±0.24	1.02±0.19	2.01±0.38
SAC + VAE	1.13±0.14	5.30±0.68	10.00±0.00
SAC + SB	0.98±0.88	2.08±0.30	10.00±0.00
DreamerV3	**0.42±0.06**	**1.23±0.20**	**1.44±0.58**
DreamerV3 w/o dec.	1.13±0.12	1.79±0.61	2.57±0.91

Results: Performance of Variants

- **Pretraining** ID & OOD improves PPO performance.

Table 1. **Million steps to reach 80% success rate across pool sizes.** Lower is better. Best performing variant for each algorithm and pool size is highlighted in bold.

Agent	Pool 1	Pool 5	Pool 10
PPO	1.75±0.44	7.80±1.08	9.73±0.36
PPO + PT (ID)	**0.95±0.21**	**5.55±1.22**	**9.17±1.10**
PPO + PT (OOD)	1.34±0.42	7.03±1.07	9.70±0.41
SAC	0.33±0.07	0.91±0.12	2.03±0.38
SAC + RAD	**0.24±0.03**	**0.42±0.06**	**0.82±0.18**
SAC + CURL	0.46±0.10	1.56±0.31	5.24±1.92
SAC + SPR	2.09±0.81	3.68±1.68	10.00±0.00
SAC + DBC	0.99±0.25	1.12±0.22	2.13±0.41
SAC + AE	1.04±0.24	1.02±0.19	2.01±0.38
SAC + VAE	1.13±0.14	5.30±0.68	10.00±0.00
SAC + SB	0.98±0.88	2.08±0.30	10.00±0.00
DreamerV3	**0.42±0.06**	**1.23±0.20**	**1.44±0.58**
DreamerV3 w/o dec.	1.13±0.12	1.79±0.61	2.57±0.91

Results: Performance of Variants

- **Pretraining** ID & OOD improves PPO performance.
- **Decoding** helps DreamerV3.

Table 1. **Million steps to reach 80% success rate across pool sizes.** Lower is better. Best performing variant for each algorithm and pool size is highlighted in bold.

Agent	Pool 1	Pool 5	Pool 10
PPO	1.75±0.44	7.80±1.08	9.73±0.36
PPO + PT (ID)	**0.95±0.21**	**5.55±1.22**	**9.17±1.10**
PPO + PT (OOD)	1.34±0.42	7.03±1.07	9.70±0.41
SAC	0.33±0.07	0.91±0.12	2.03±0.38
SAC + RAD	**0.24±0.03**	**0.42±0.06**	**0.82±0.18**
SAC + CURL	0.46±0.10	1.56±0.31	5.24±1.92
SAC + SPR	2.09±0.81	3.68±1.68	10.00±0.00
SAC + DBC	0.99±0.25	1.12±0.22	2.13±0.41
SAC + AE	1.04±0.24	1.02±0.19	2.01±0.38
SAC + VAE	1.13±0.14	5.30±0.68	10.00±0.00
SAC + SB	0.98±0.88	2.08±0.30	10.00±0.00
DreamerV3	**0.42±0.06**	**1.23±0.20**	**1.44±0.58**
DreamerV3 w/o dec.	1.13±0.12	1.79±0.61	2.57±0.91

Results: Performance of Variants

- **Pretraining** ID & OOD improves PPO performance.
- **Decoding** helps DreamerV3.
- SAC with **Data Augmentation** (RAD) is highly effective.

Table 1. **Million steps to reach 80% success rate across pool sizes.** Lower is better. Best performing variant for each algorithm and pool size is highlighted in bold.

Agent	Pool 1	Pool 5	Pool 10
PPO	1.75±0.44	7.80±1.08	9.73±0.36
PPO + PT (ID)	**0.95±0.21**	**5.55±1.22**	**9.17±1.10**
PPO + PT (OOD)	1.34±0.42	7.03±1.07	9.70±0.41
SAC	0.33±0.07	0.91±0.12	2.03±0.38
SAC + RAD	**0.24±0.03**	**0.42±0.06**	**0.82±0.18**
SAC + CURL	0.46±0.10	1.56±0.31	5.24±1.92
SAC + SPR	2.09±0.81	3.68±1.68	10.00±0.00
SAC + DBC	0.99±0.25	1.12±0.22	2.13±0.41
SAC + AE	1.04±0.24	1.02±0.19	2.01±0.38
SAC + VAE	1.13±0.14	5.30±0.68	10.00±0.00
SAC + SB	0.98±0.88	2.08±0.30	10.00±0.00
DreamerV3	**0.42±0.06**	**1.23±0.20**	**1.44±0.58**
DreamerV3 w/o dec.	1.13±0.12	1.79±0.61	2.57±0.91

Results: Performance of Variants

- **Pretraining** ID & OOD improves PPO performance.
- **Decoding** helps DreamerV3.
- SAC with **Data Augmentation** (RAD) is highly effective.
- Auxiliary methods **underperform** baselines. Their **assumptions** don't seem to hold in SPGym.

Table 1. **Million steps to reach 80% success rate across pool sizes.** Lower is better. Best performing variant for each algorithm and pool size is highlighted in bold.

Agent	Pool 1	Pool 5	Pool 10
PPO	$1.75_{\pm 0.44}$	$7.80_{\pm 1.08}$	$9.73_{\pm 0.36}$
PPO + PT (ID)	$\mathbf{0.95_{\pm 0.21}}$	$\mathbf{5.55_{\pm 1.22}}$	$\mathbf{9.17_{\pm 1.10}}$
PPO + PT (OOD)	$1.34_{\pm 0.42}$	$7.03_{\pm 1.07}$	$9.70_{\pm 0.41}$
SAC	$0.33_{\pm 0.07}$	$0.91_{\pm 0.12}$	$2.03_{\pm 0.38}$
SAC + RAD	$\mathbf{0.24_{\pm 0.03}}$	$\mathbf{0.42_{\pm 0.06}}$	$\mathbf{0.82_{\pm 0.18}}$
SAC + CURL	$0.46_{\pm 0.10}$	$1.56_{\pm 0.31}$	$5.24_{\pm 1.92}$
SAC + SPR	$2.09_{\pm 0.81}$	$3.68_{\pm 1.68}$	$10.00_{\pm 0.00}$
SAC + DBC	$0.99_{\pm 0.25}$	$1.12_{\pm 0.22}$	$2.13_{\pm 0.41}$
SAC + AE	$1.04_{\pm 0.24}$	$1.02_{\pm 0.19}$	$2.01_{\pm 0.38}$
SAC + VAE	$1.13_{\pm 0.14}$	$5.30_{\pm 0.68}$	$10.00_{\pm 0.00}$
SAC + SB	$0.98_{\pm 0.88}$	$2.08_{\pm 0.30}$	$10.00_{\pm 0.00}$
DreamerV3	$\mathbf{0.42_{\pm 0.06}}$	$\mathbf{1.23_{\pm 0.20}}$	$\mathbf{1.44_{\pm 0.58}}$
DreamerV3 w/o dec.	$1.13_{\pm 0.12}$	$1.79_{\pm 0.61}$	$2.57_{\pm 0.91}$

Conclusion

SPGym is a new benchmark that **isolates the visual representation challenge** from the environment dynamics, rewards, state and action spaces.

Conclusion

SPGym is a new benchmark that **isolates the visual representation challenge** from the environment dynamics, rewards, state and action spaces.

- Sophisticated representation learning techniques **struggle** with SPGym's unique characteristics.

Conclusion

SPGym is a new benchmark that **isolates the visual representation challenge** from the environment dynamics, rewards, state and action spaces.

- Sophisticated representation learning techniques **struggle** with SPGym's unique characteristics.
- Agents seem to **memorize** specific visual features rather than understand the underlying task structure.

Conclusion

SPGym is a new benchmark that **isolates the visual representation challenge** from the environment dynamics, rewards, state and action spaces.

- Sophisticated representation learning techniques **struggle** with SPGym's unique characteristics.
- Agents seem to **memorize** specific visual features rather than understand the underlying task structure.
- Simply increasing the **diversity of training data is not enough** to bridge this gap with current algorithms.

Thank You!

Bryan de Oliveira[1,2], Luana G. B. Martins[1], Bruno Brandão[1,2], Murilo L. da Luz[1,2], Telma W. de L. Soares[1,2], Luckeciano C. Melo[1,3]

[1]Advanced Knowledge Center for Immersive Technologies (AKCIT)
[2]Institute of Informatics, Federal University of Goiás
[3]OATML, University of Oxford

Correspondence to: bryanlincoln@discente.ufg.br
Code: https://github.com/bryanoliveira/sliding-puzzles-gym