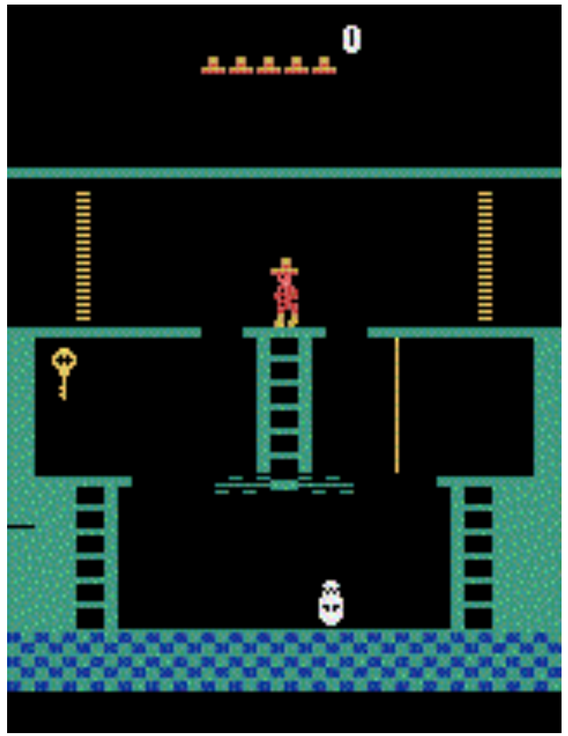
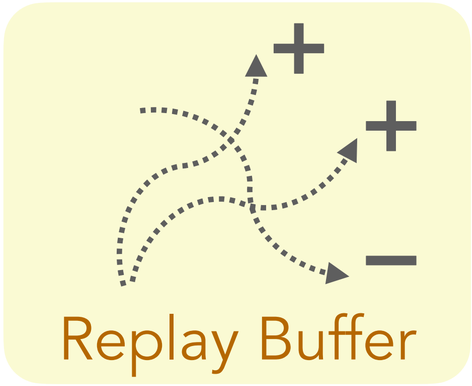
Leveraging Skills from Unlabeled Prior Data for Efficient Online Exploration

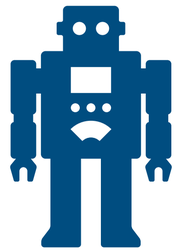
Max Wilcoxson*, Qiyang (Colin) Li*, Kevin Frans, Sergey Levine
UC Berkeley

Exploration is Hard



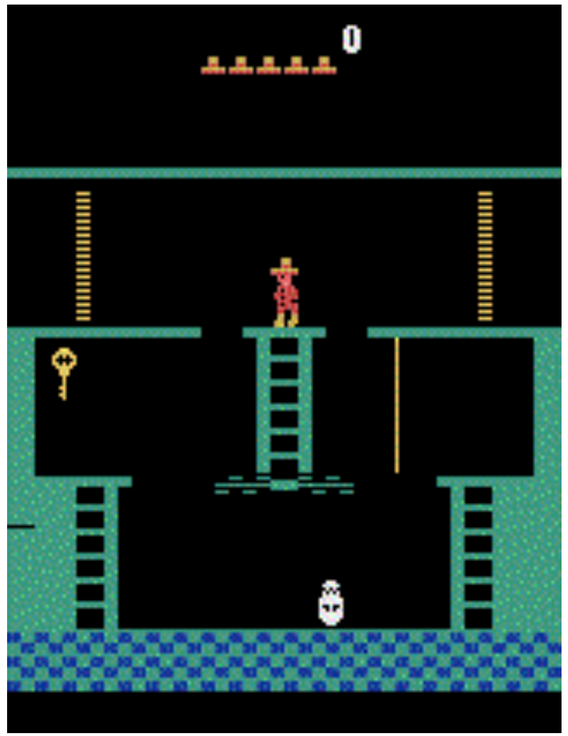
Online



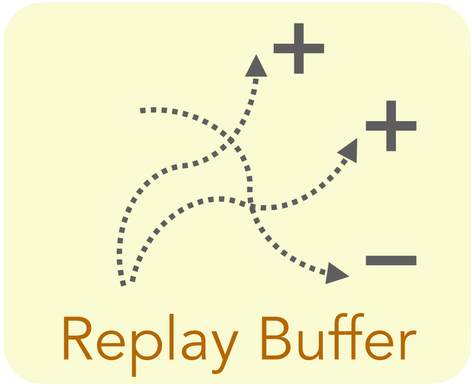


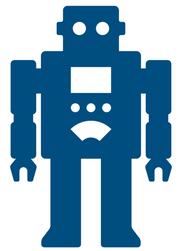
https://www.gymlibrary.dev/environments/atari/montezuma_revenge/

Exploration is Hard



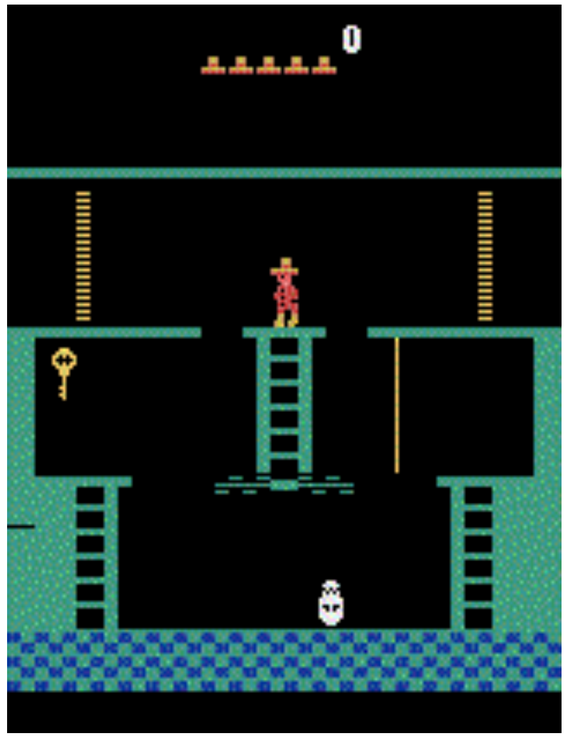
Online





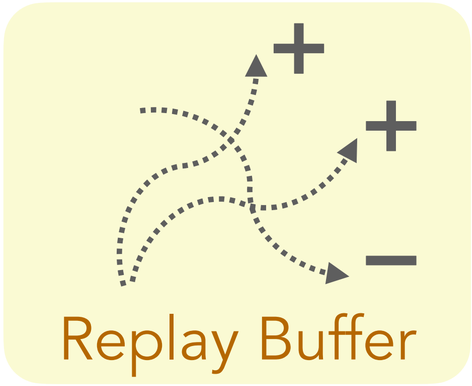
https://www.gymlibrary.dev/environments/atari/montezuma_revenge/

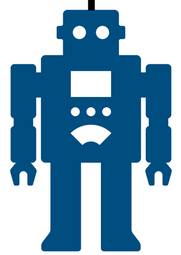
Exploration is Hard



Online

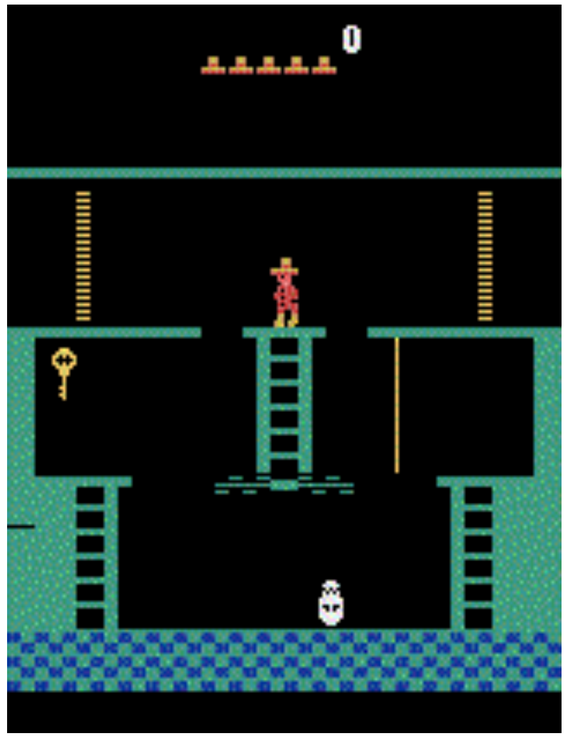
Explore

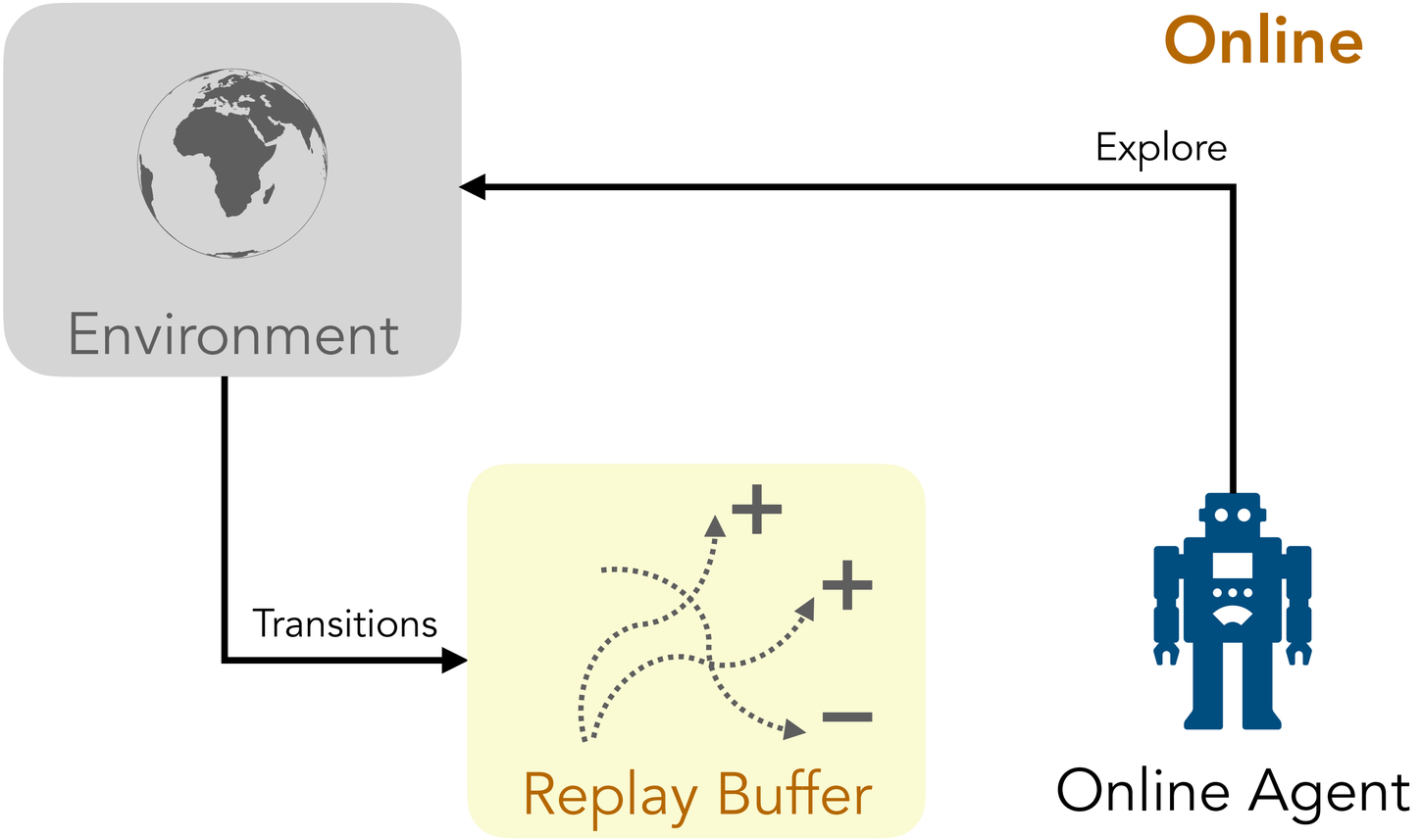




https://www.gymlibrary.dev/environments/atari/montezuma_revenge/

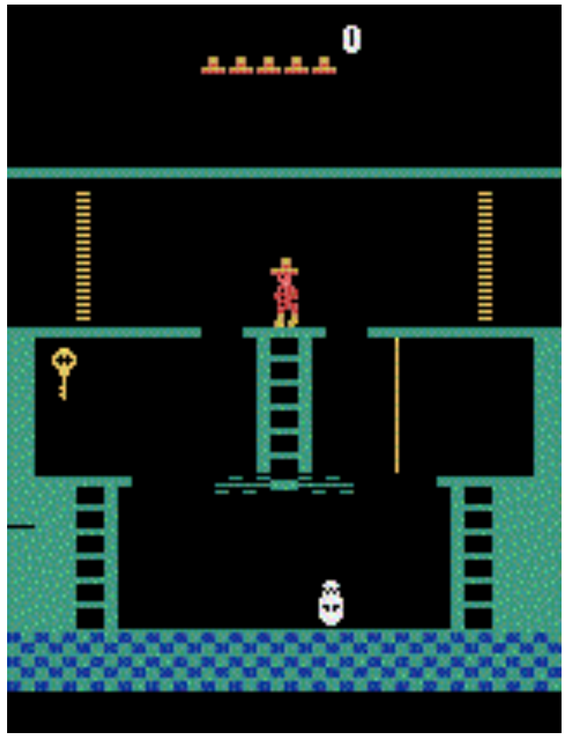
Exploration is Hard

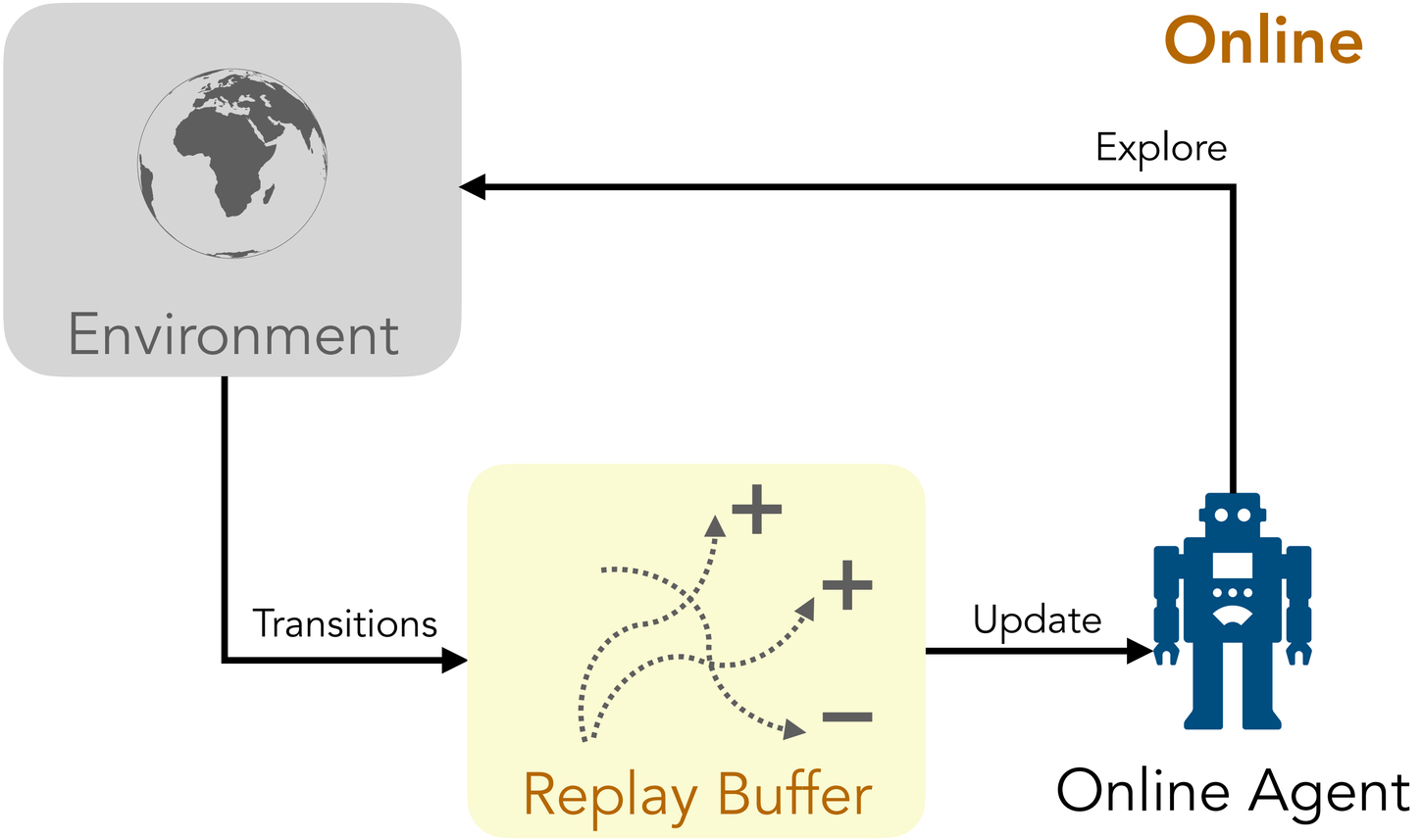




https://www.gymlibrary.dev/environments/atari/montezuma_revenge/

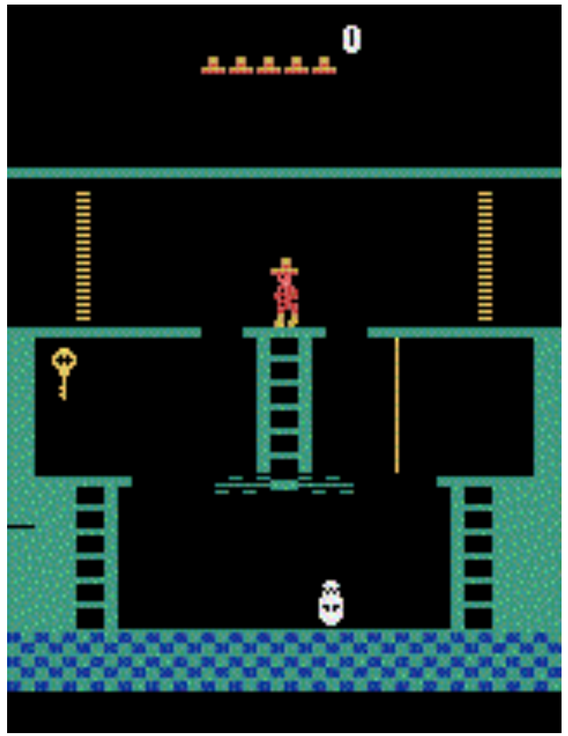
Exploration is Hard

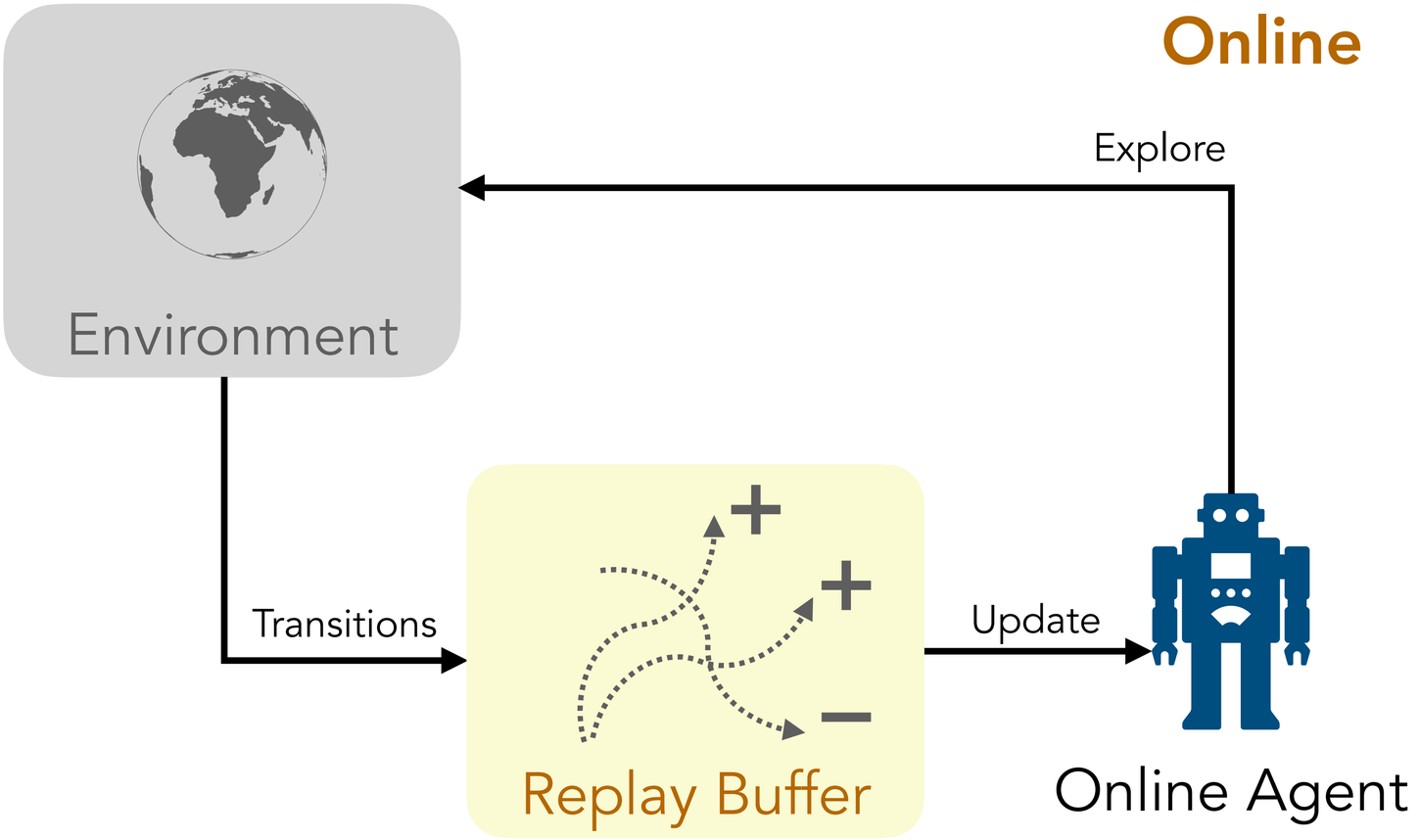




https://www.gymlibrary.dev/environments/atari/montezuma_revenge/

Exploration is Hard

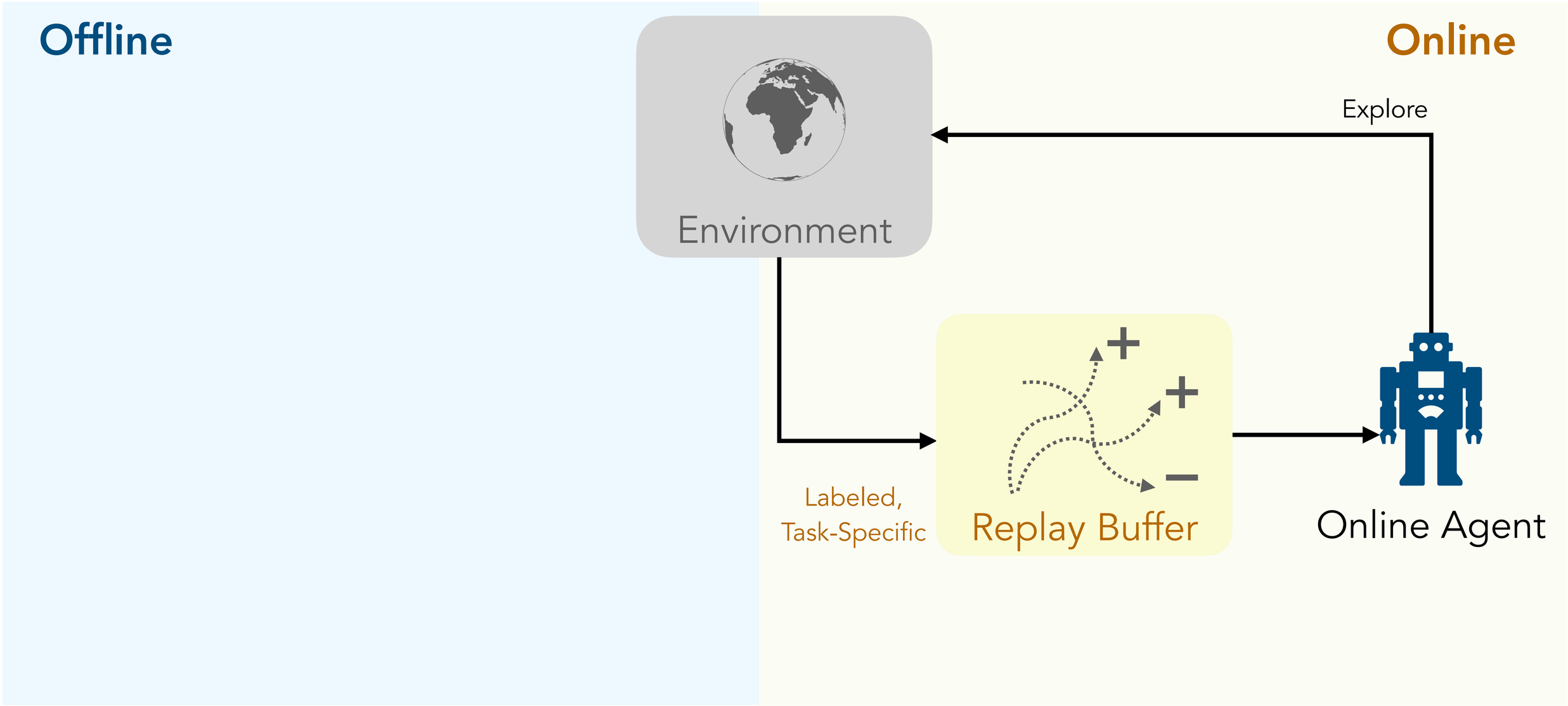




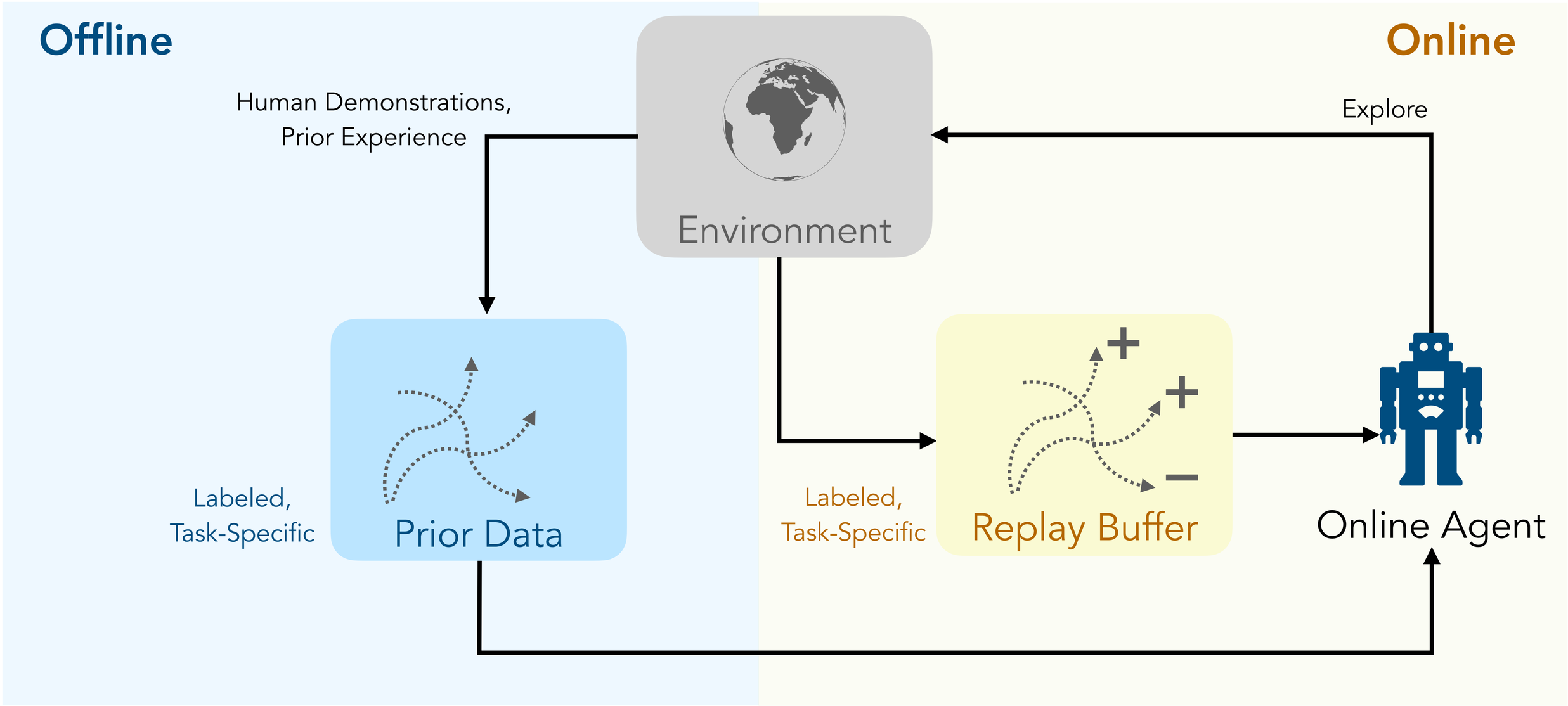
In the worst case, we must visit **every possible state** in the environment

https://www.gymlibrary.dev/environments/atari/montezuma_revenge/

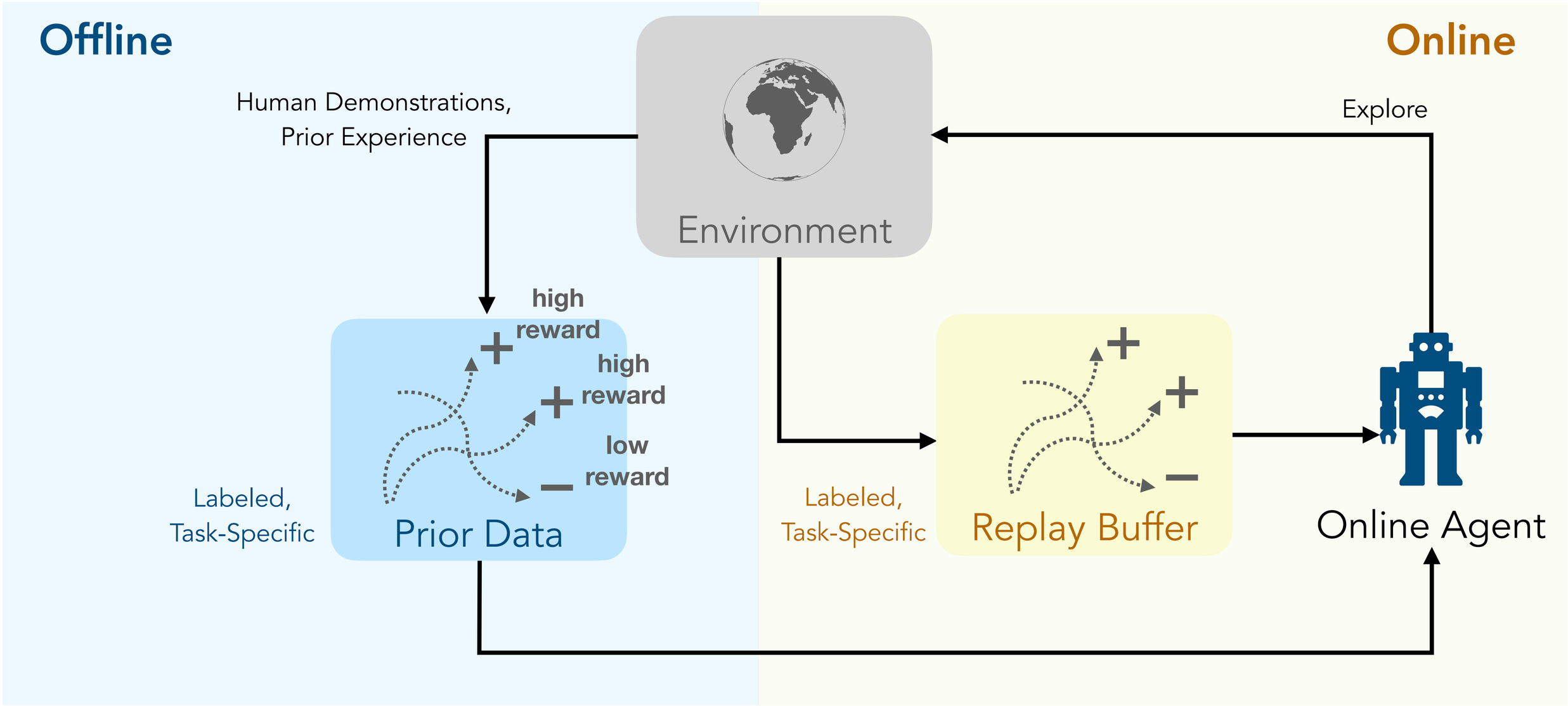
Data-driven Exploration



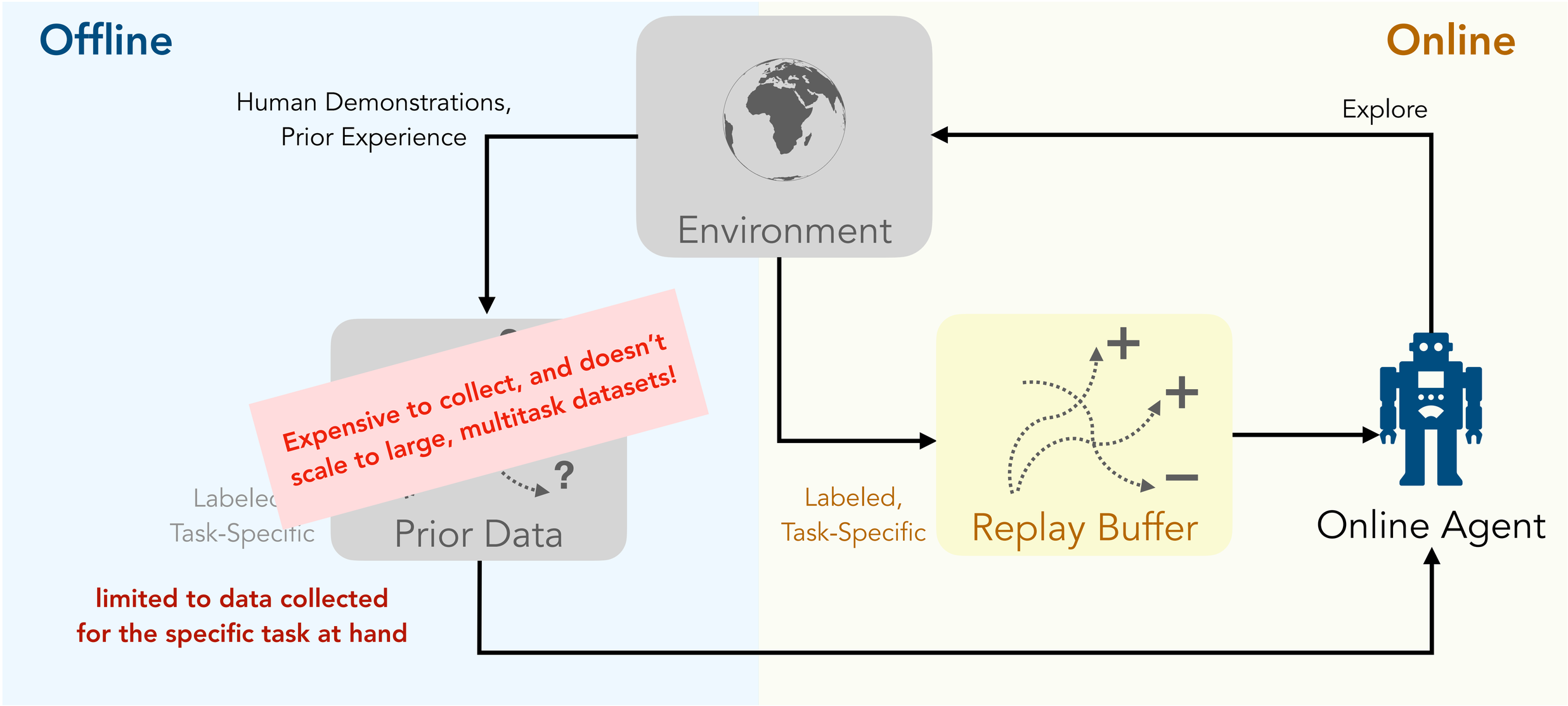
Data-driven Exploration



Data-driven Exploration

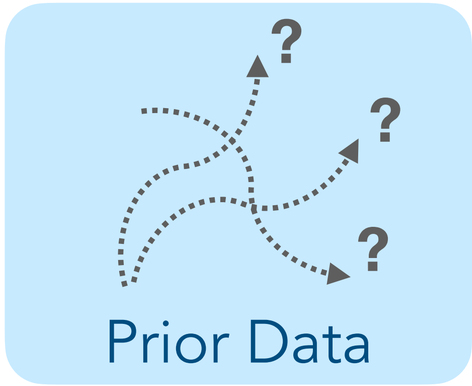


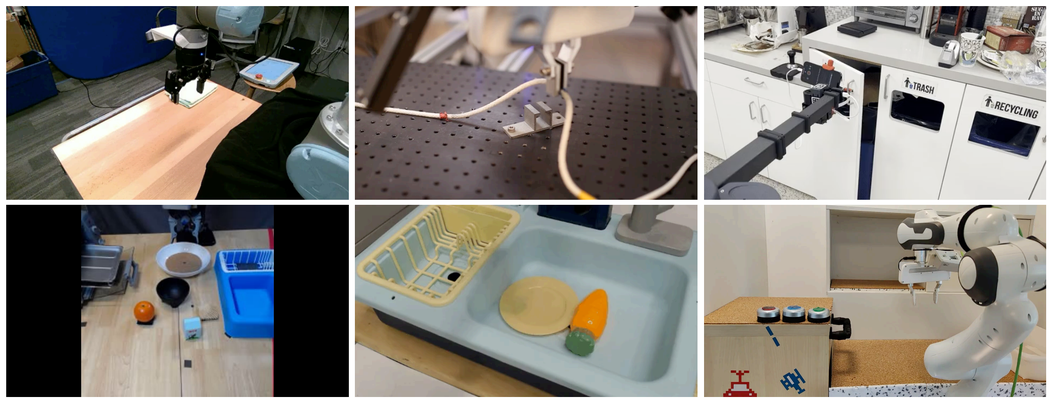
Data-driven Exploration



Data-driven Exploration

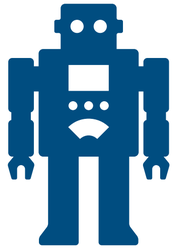
Unlabeled,
Task-Agnostic





**can leverage data collected for a range of tasks,
scales without expensive supervision**

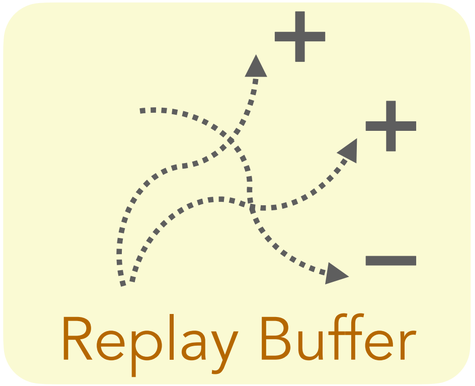
???



Online Policy

Labeled,
Task-Specific

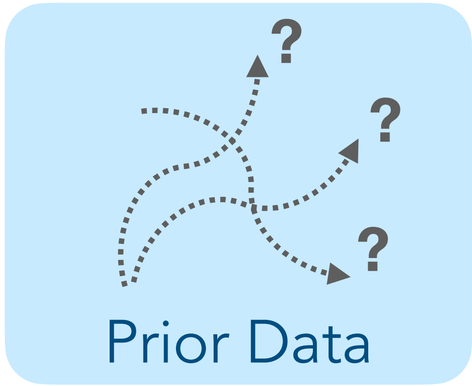
Off-policy RL

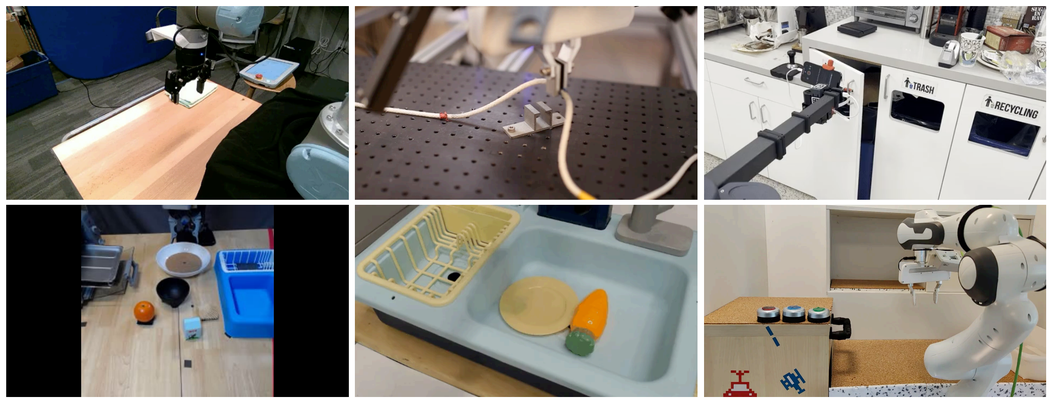


https://robotics-transformer-x.github.io/

Data-driven Exploration

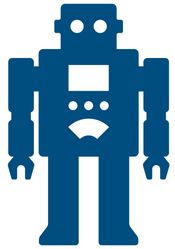
Unlabeled,
Task-Agnostic





**can leverage data collected for a range of tasks,
scales without expensive supervision**

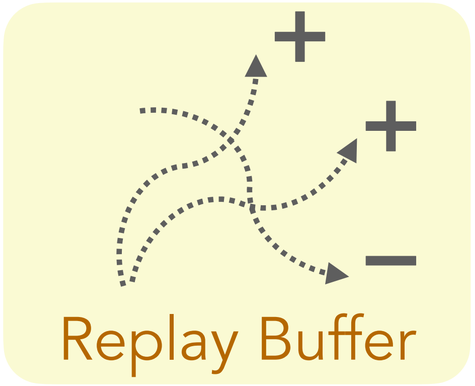
???



Online Policy

Labeled,
Task-Specific

Off-policy RL



https://robotics-transformer-x.github.io/

Data-driven Exploration

Unlabeled,
Task-Agnostic

Prior Data

???

Online Policy

How can we leverage **unlabeled** prior data for online exploration?

Labeled,
Task-Specific

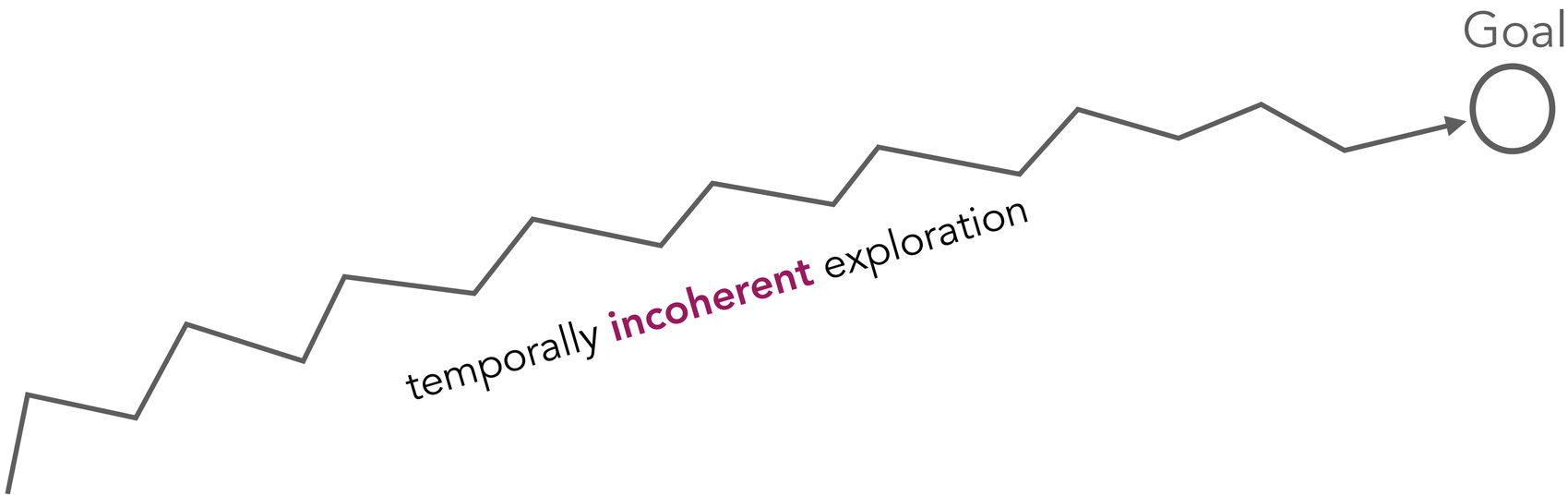
Off-policy RL

Replay Buffer

can leverage data collected for a range of tasks,
scales without expensive supervision

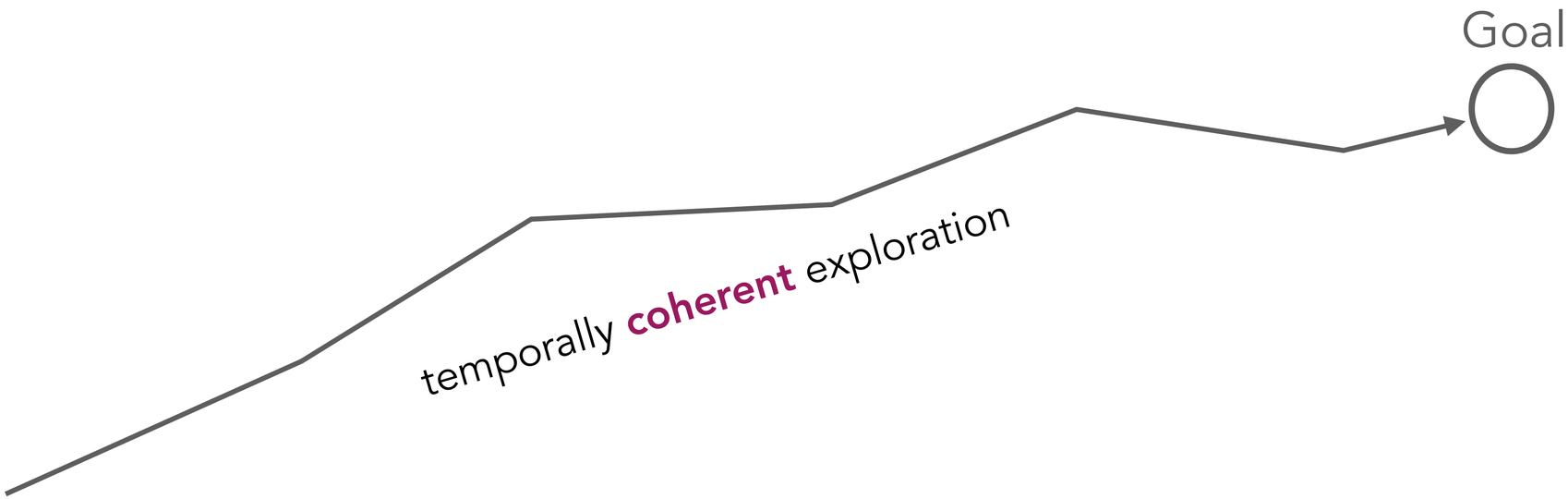
https://robotics-transformer-x.github.io/

Challenge #1: Unstructured Exploration



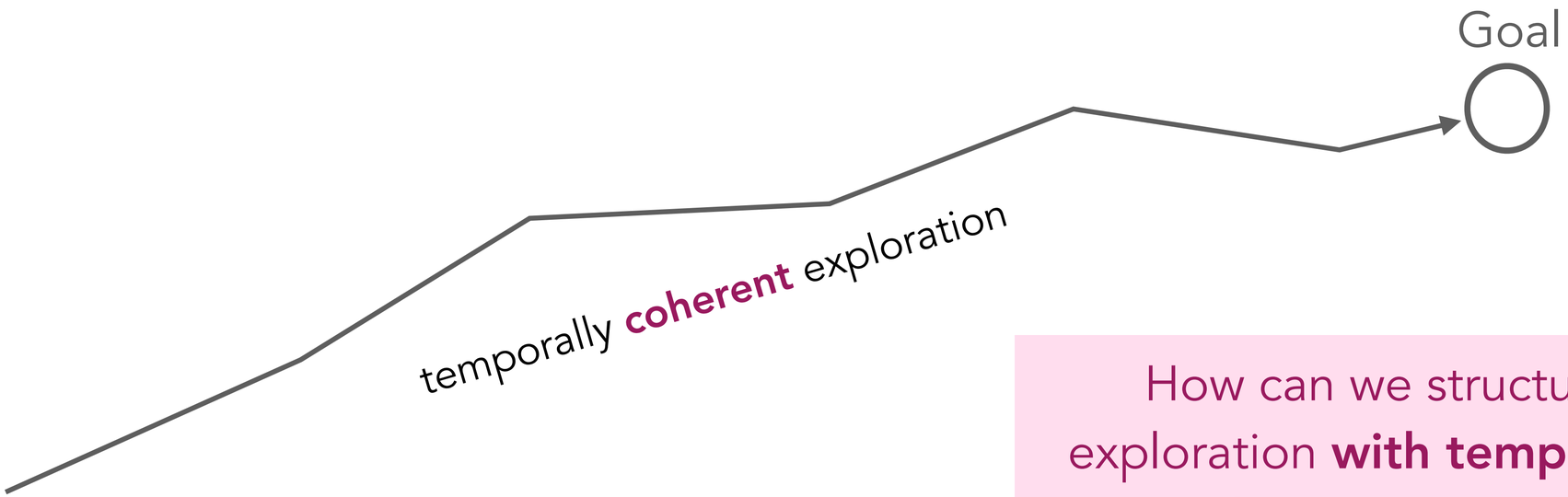
Agent actions initially lack coherency,
leading to **high sample complexity**

Challenge #1: Unstructured Exploration



Temporally coherent actions reduces task horizon,
improving sample complexity

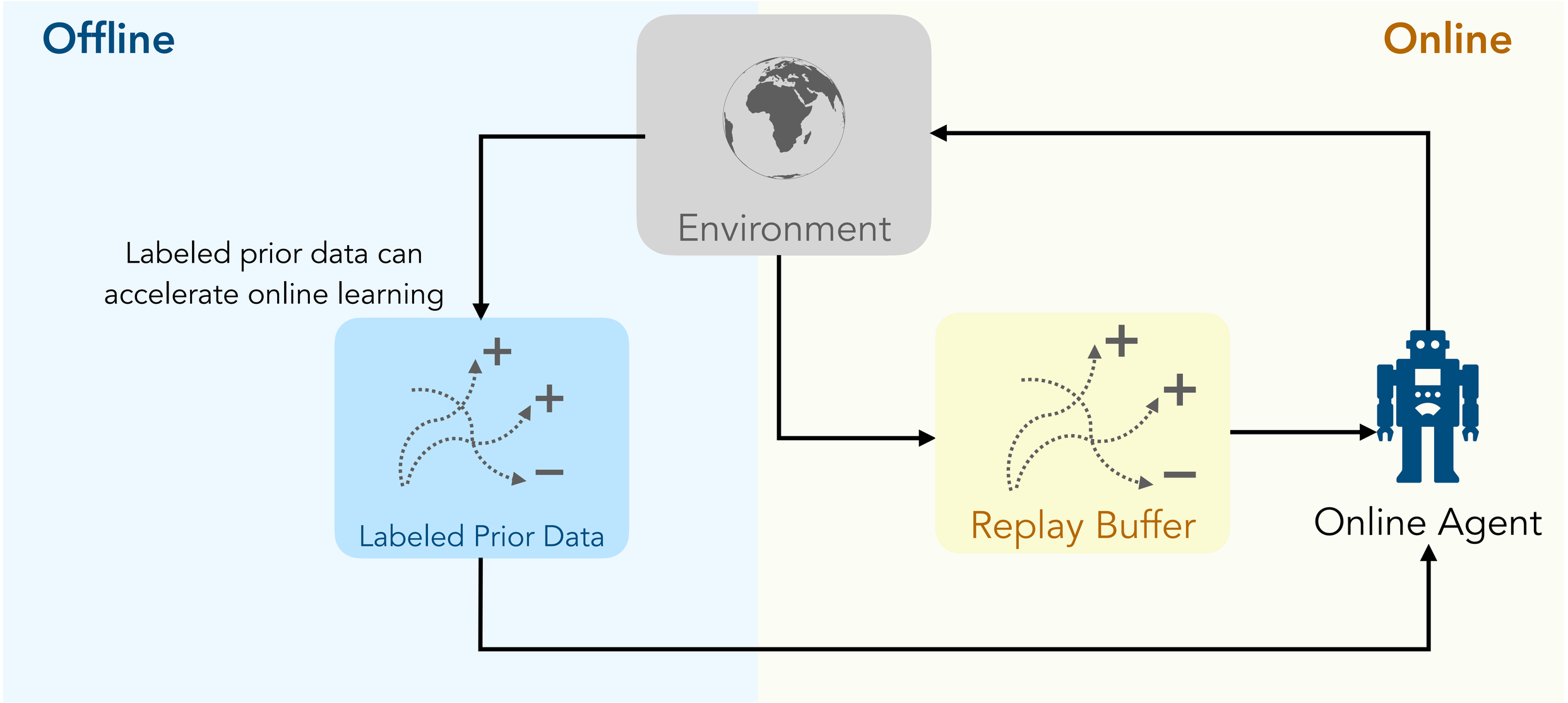
Challenge #1: Unstructured Exploration



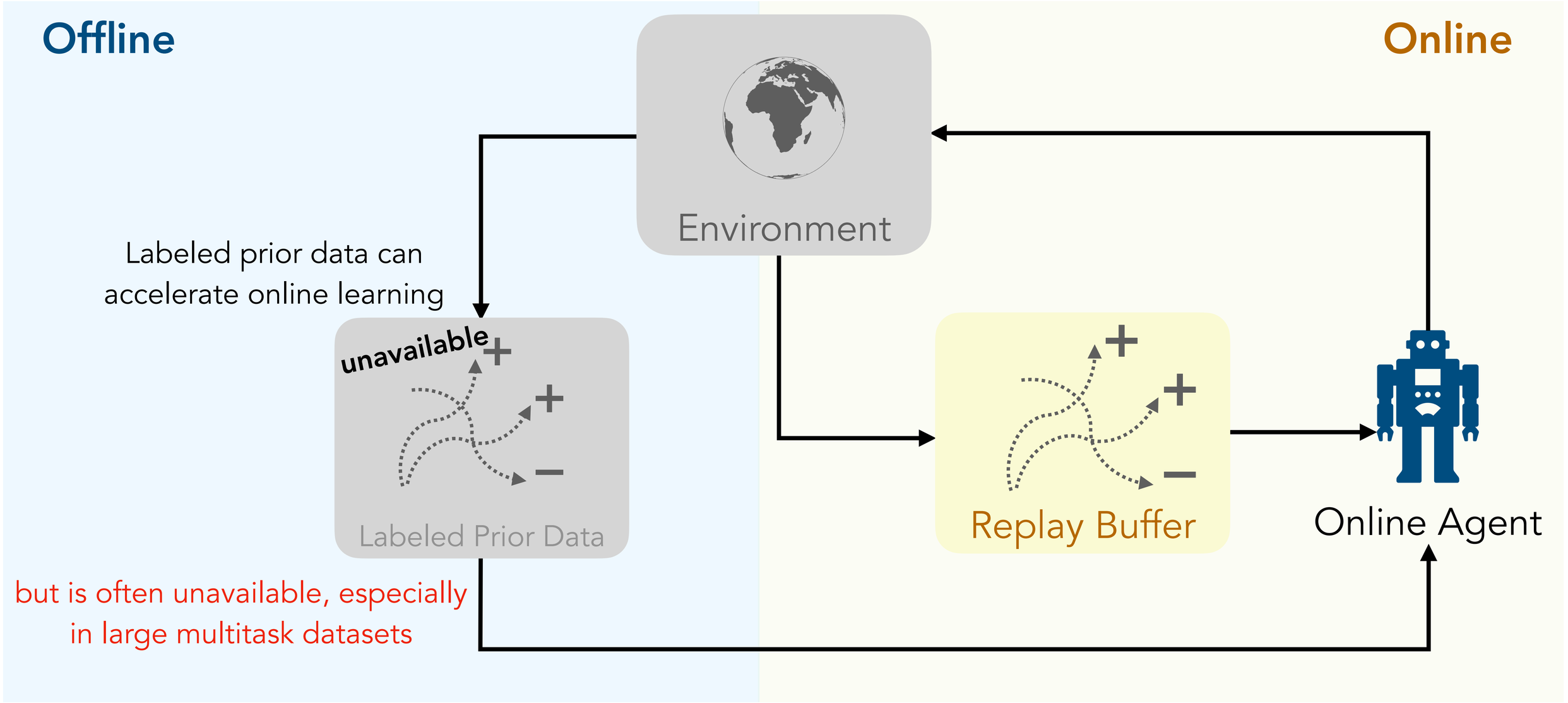
Temporally coherent actions reduces task horizon,
improving sample complexity

How can we structure exploration **with temporally coherent actions?**

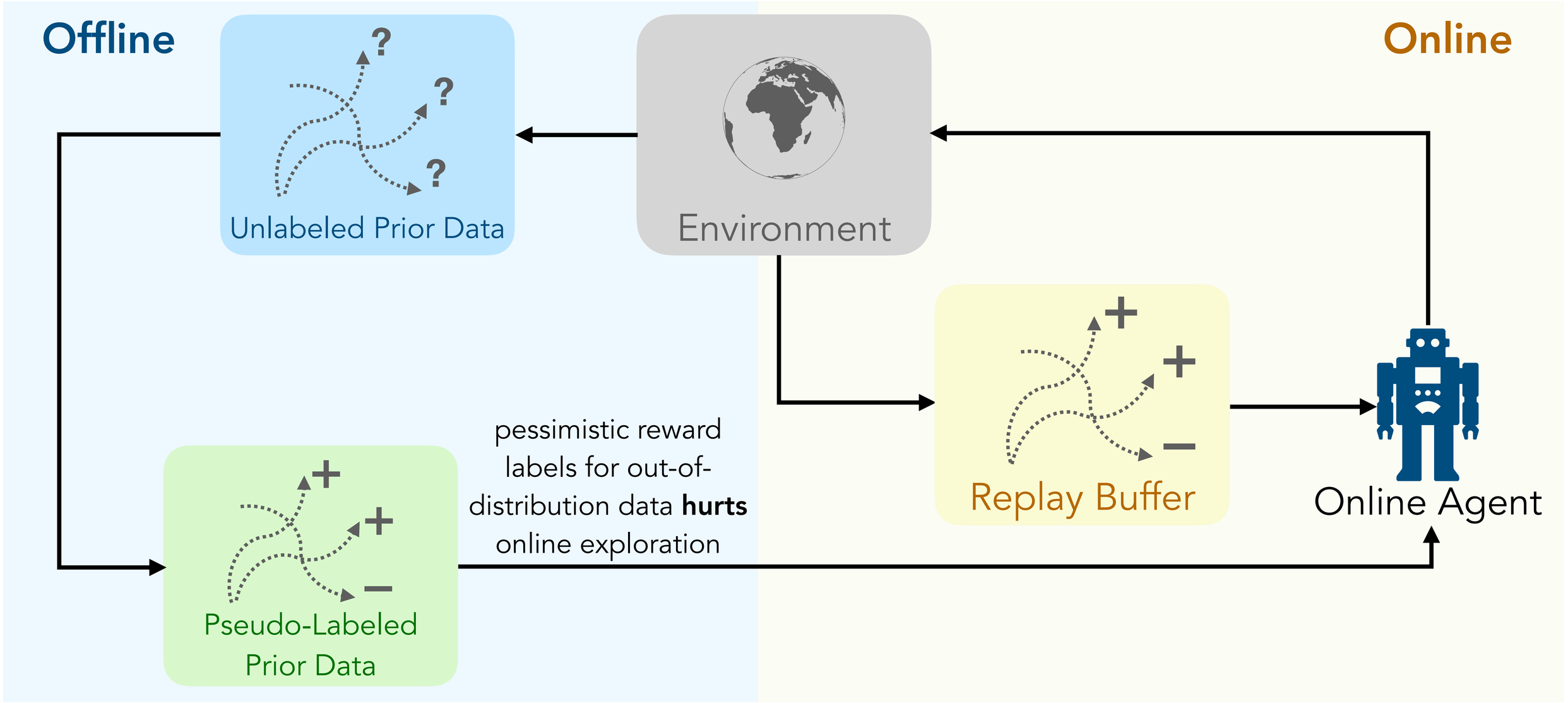
Challenge #2: No access to task-specific reward labels



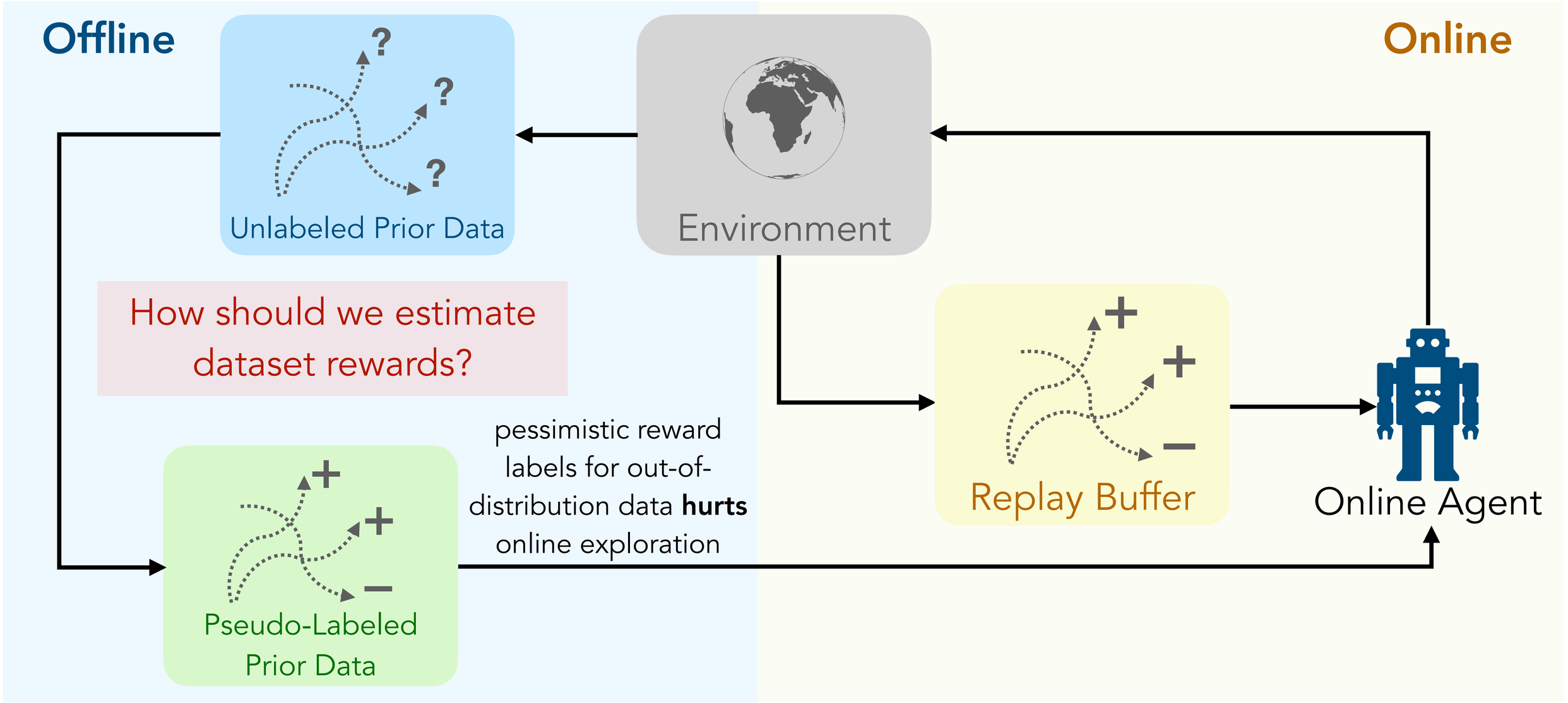
Challenge #2: No access to task-specific reward labels



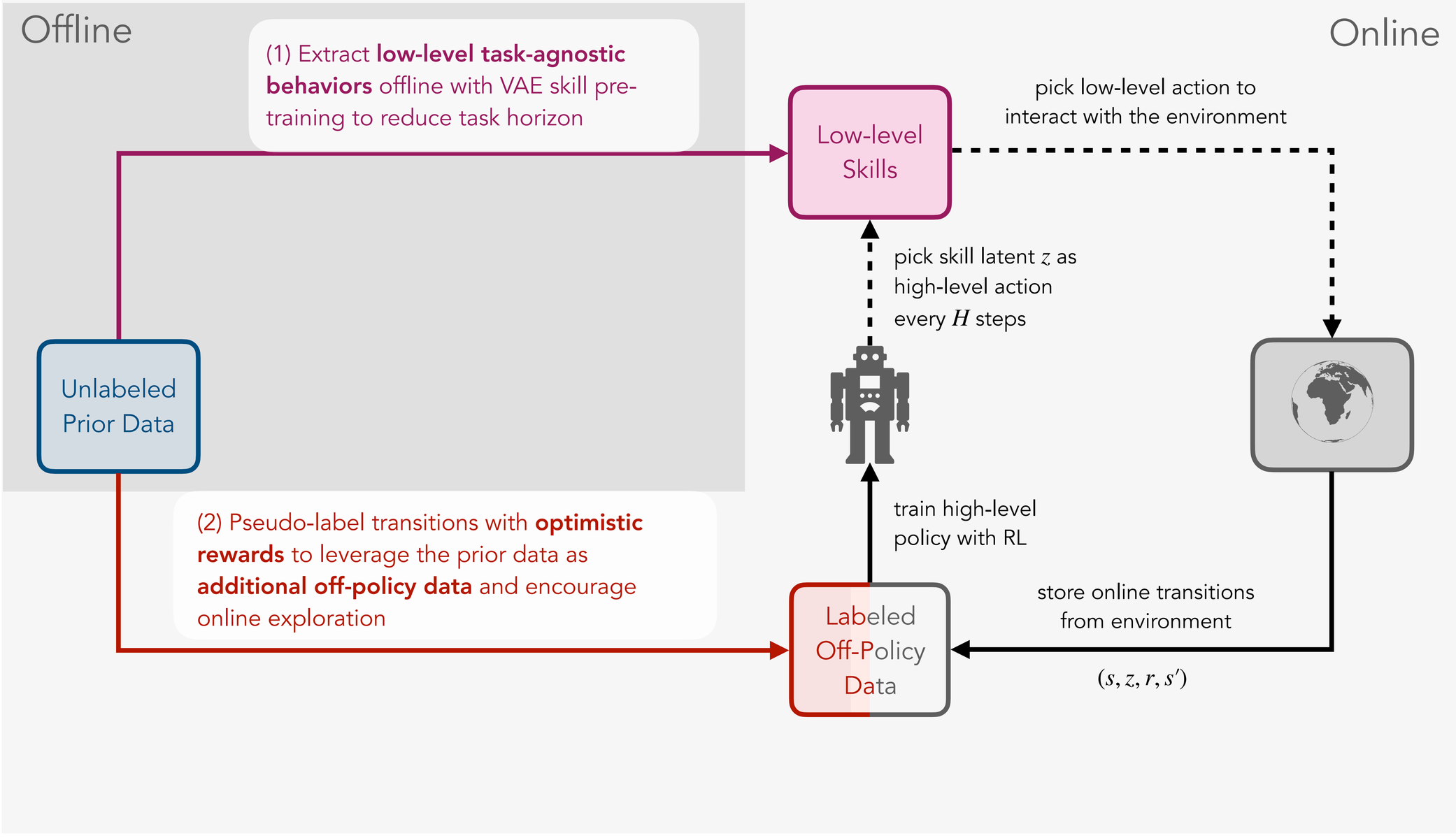
Challenge #2: No access to task-specific reward labels



Challenge #2: No access to task-specific reward labels

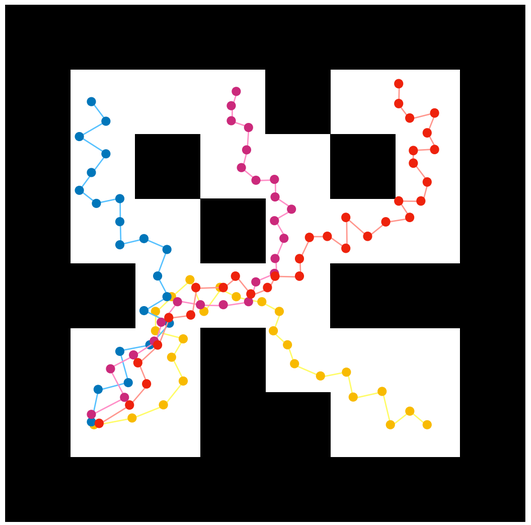


SUPE: **S**kills from **U**nlabeled **P**rior Data for **E**xploration

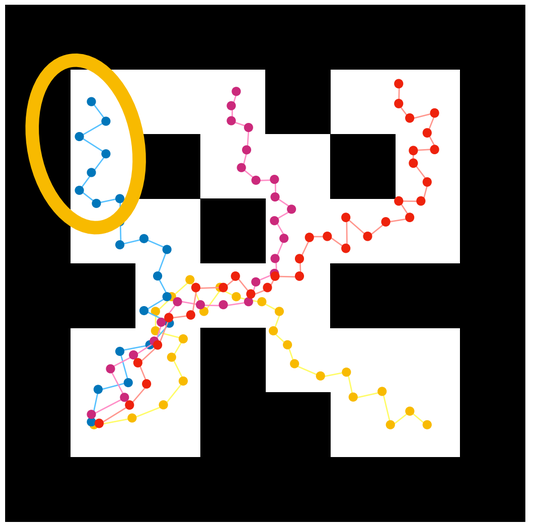


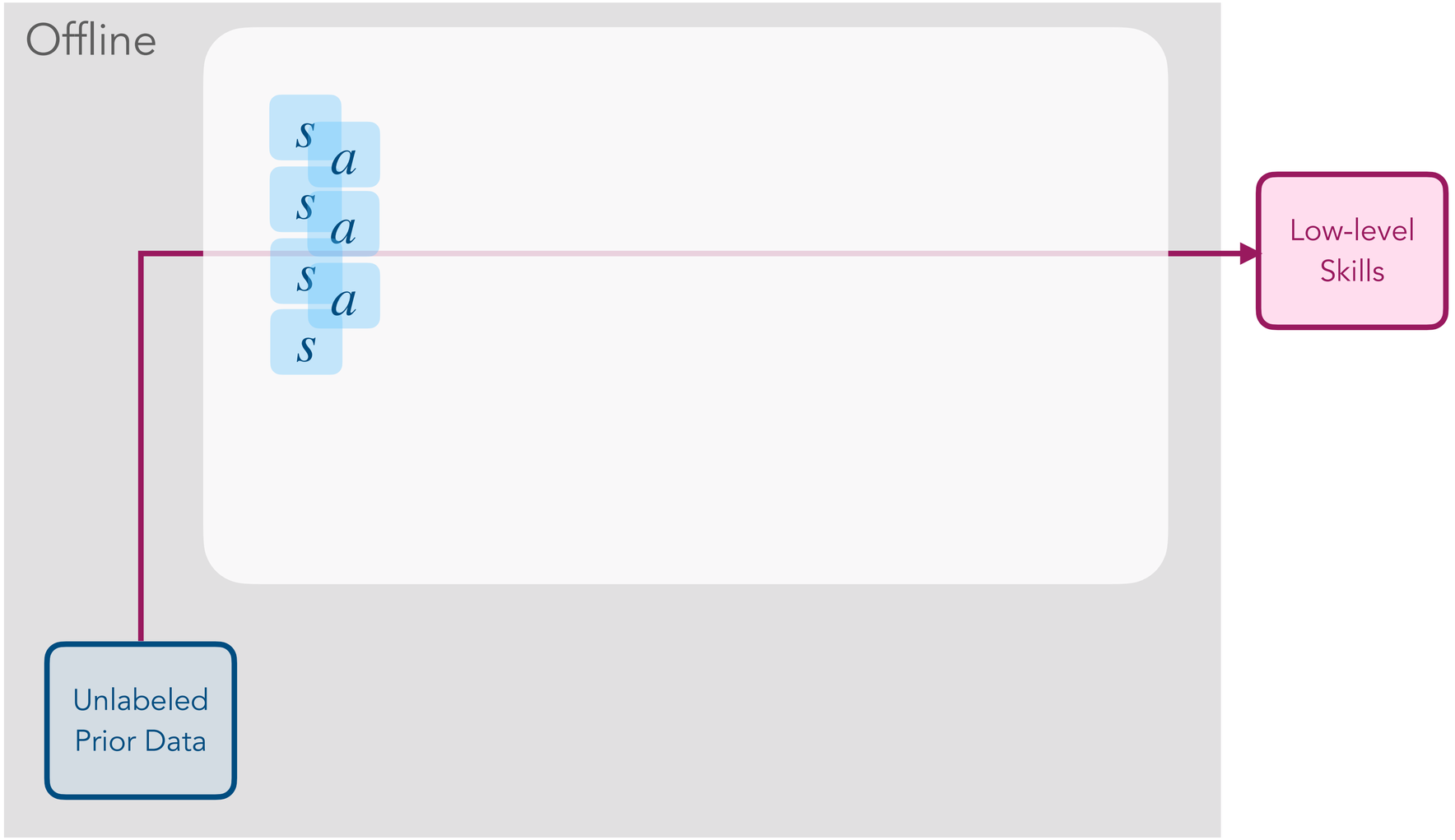
(1) Extract **low-level task-agnostic behaviors** offline
with VAE skill pre-training to reduce task horizon

(1) Extract **low-level task-agnostic behaviors** offline with VAE skill pre-training to reduce task horizon

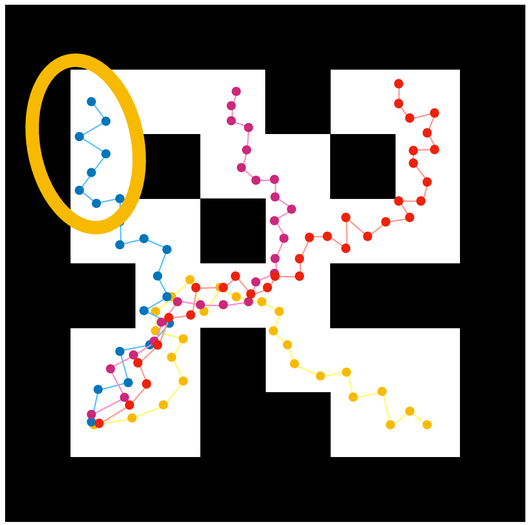


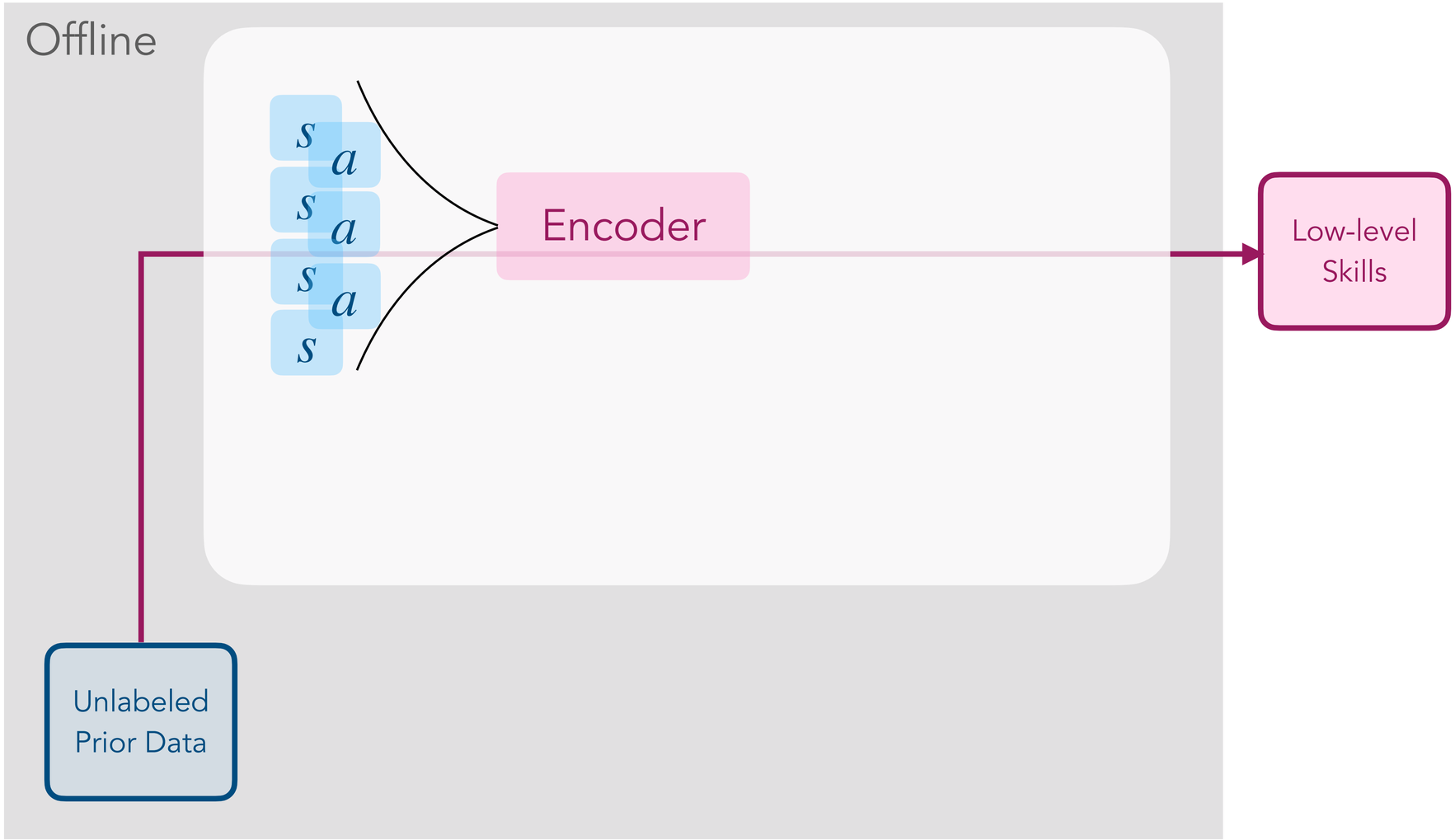
(1) Extract **low-level task-agnostic behaviors** offline
with VAE skill pre-training to reduce task horizon



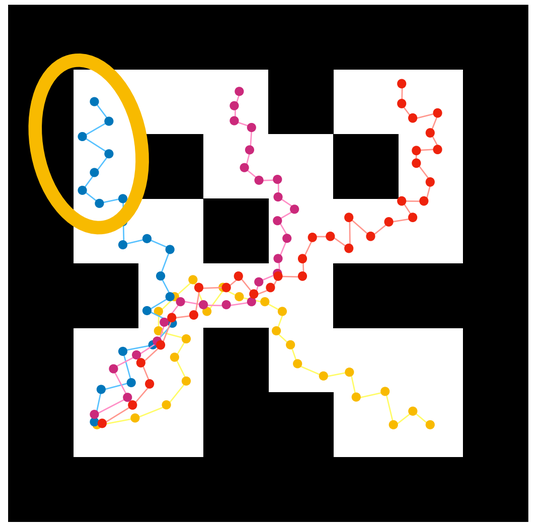


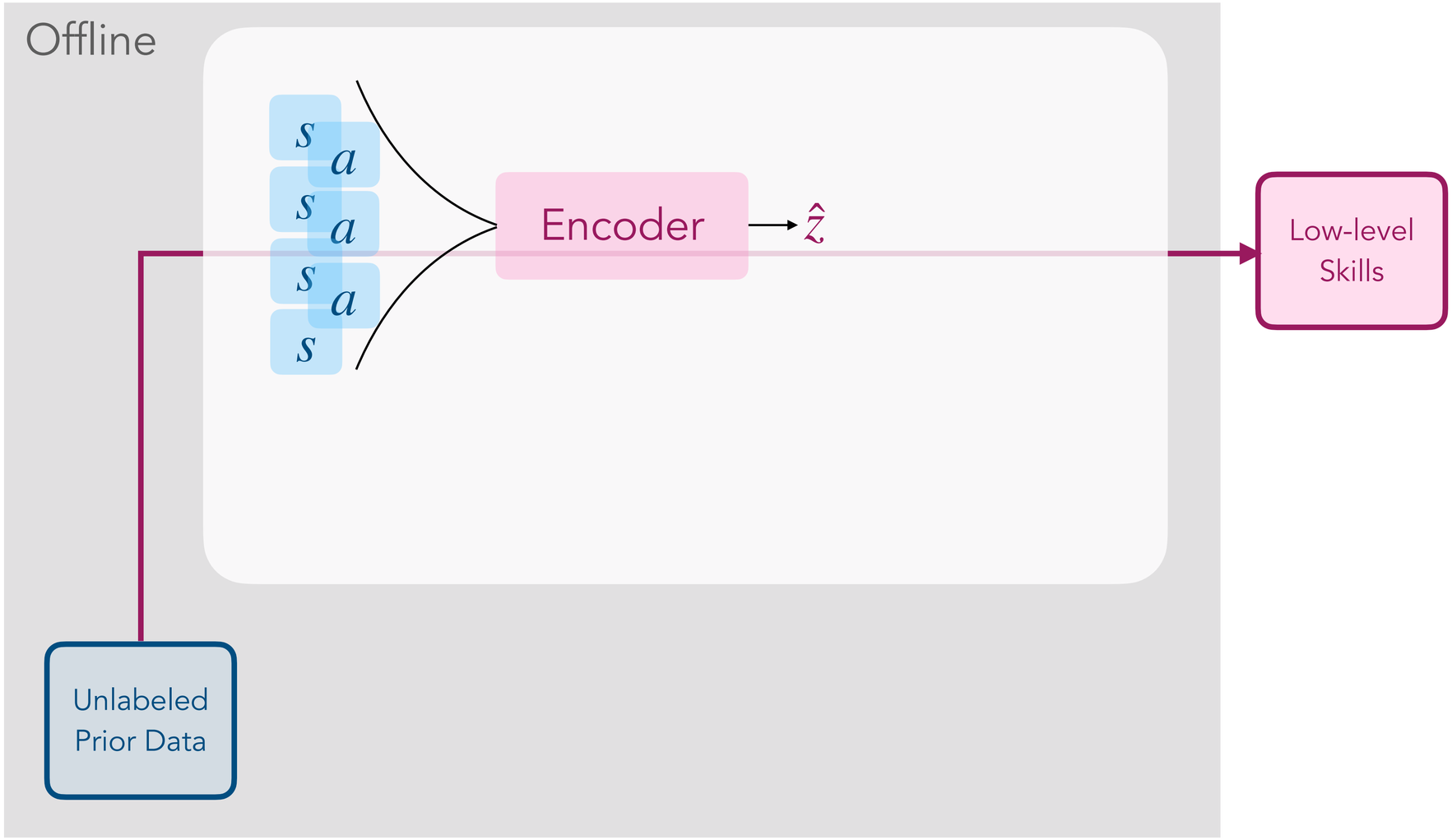
(1) Extract **low-level task-agnostic behaviors** offline with VAE skill pre-training to reduce task horizon



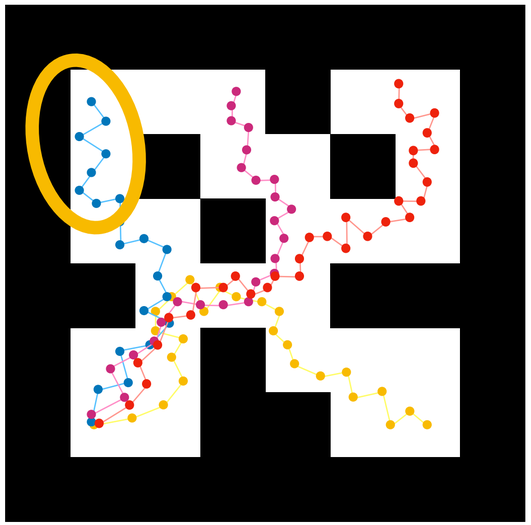


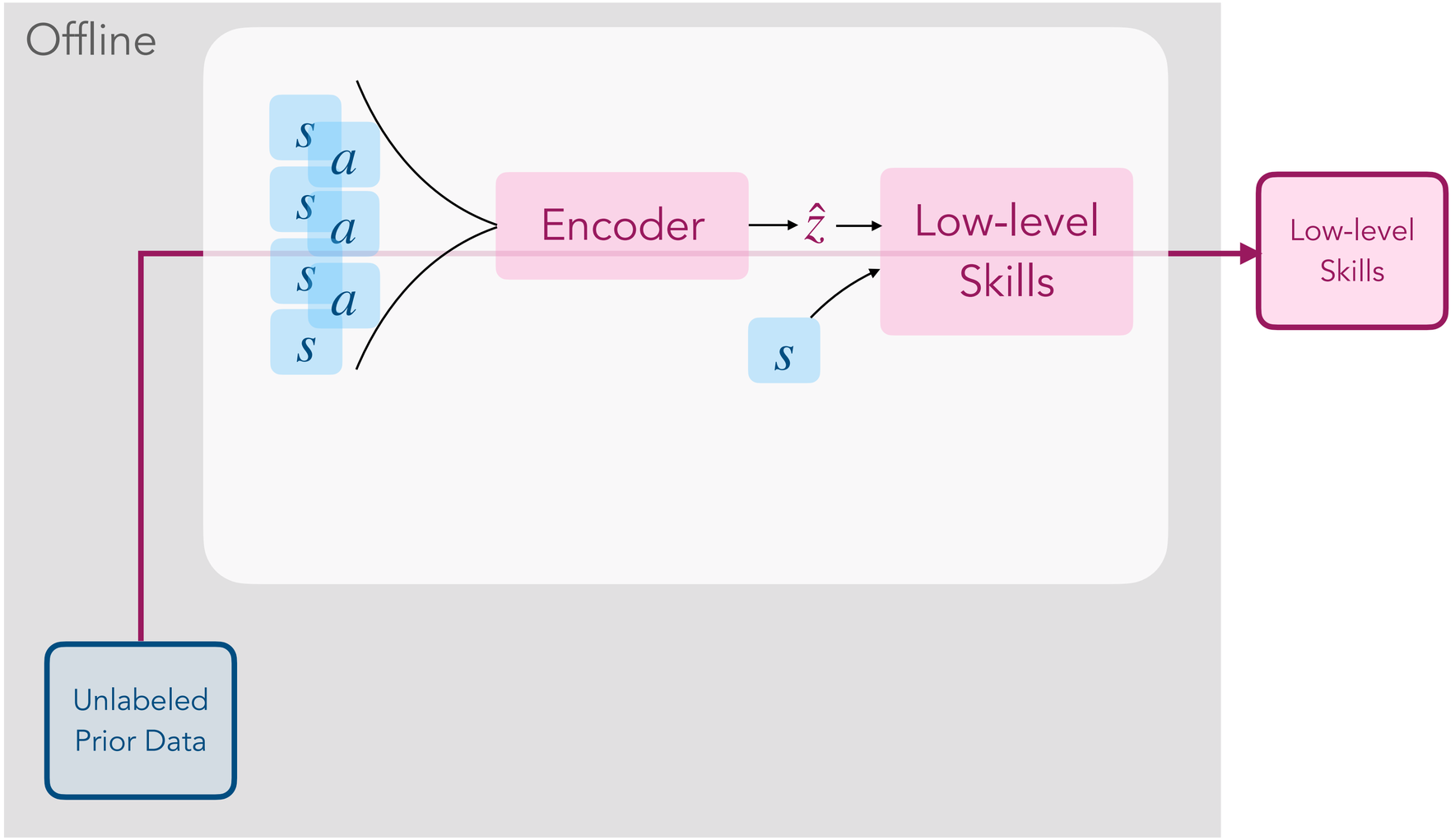
(1) Extract **low-level task-agnostic behaviors** offline with VAE skill pre-training to reduce task horizon



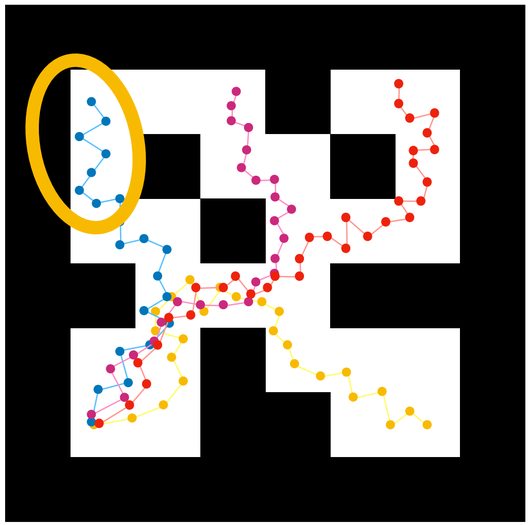


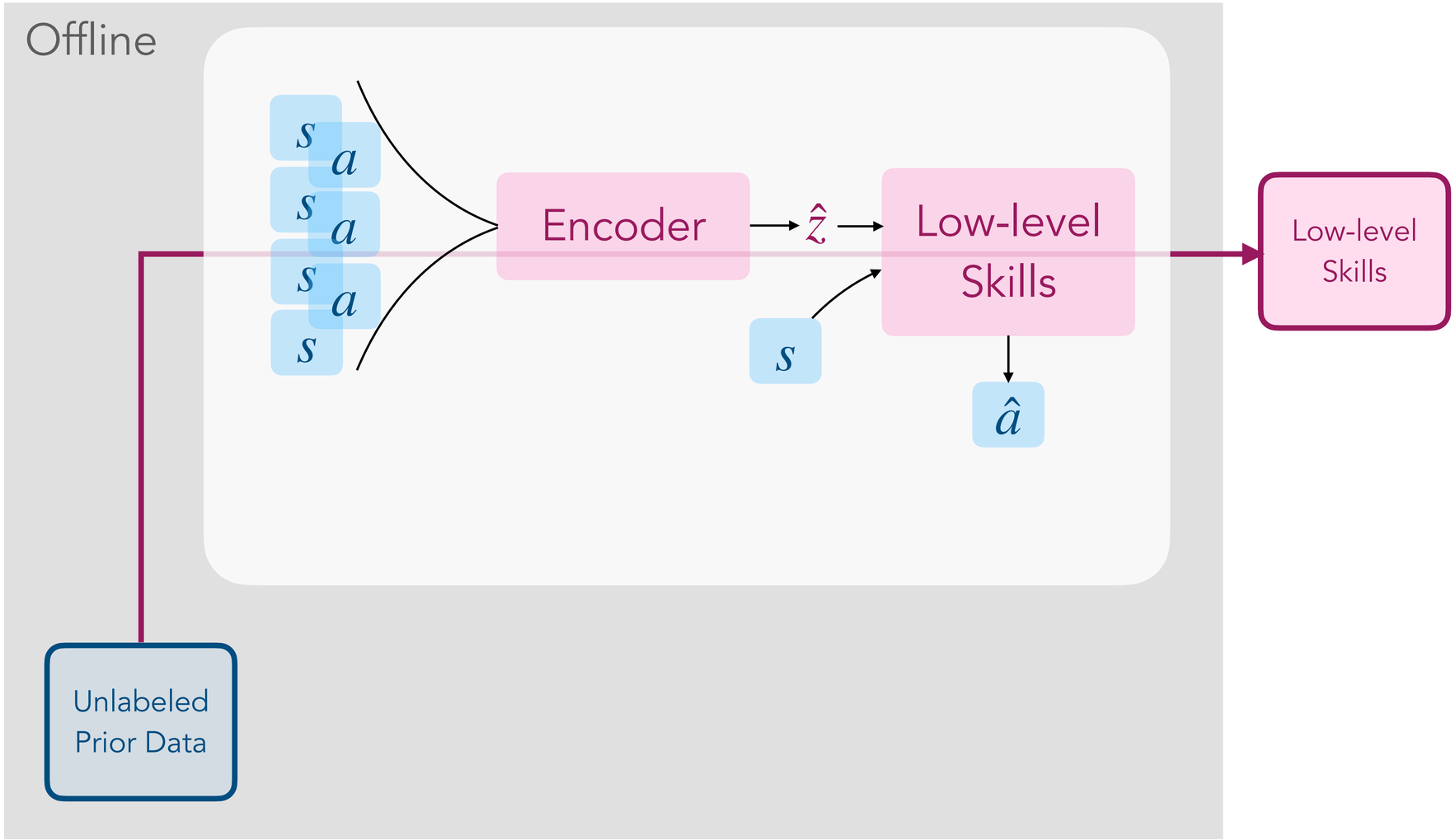
(1) Extract **low-level task-agnostic behaviors** offline
with VAE skill pre-training to reduce task horizon



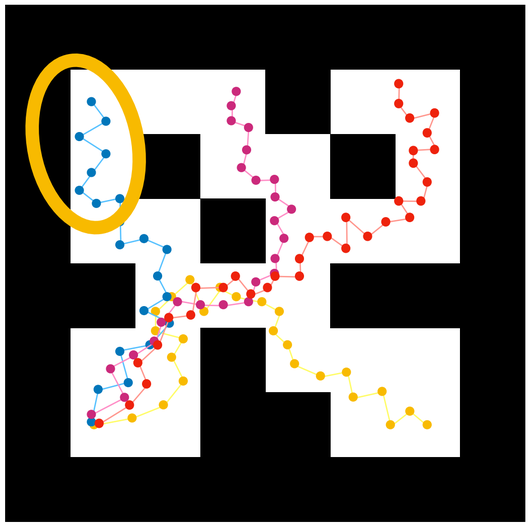


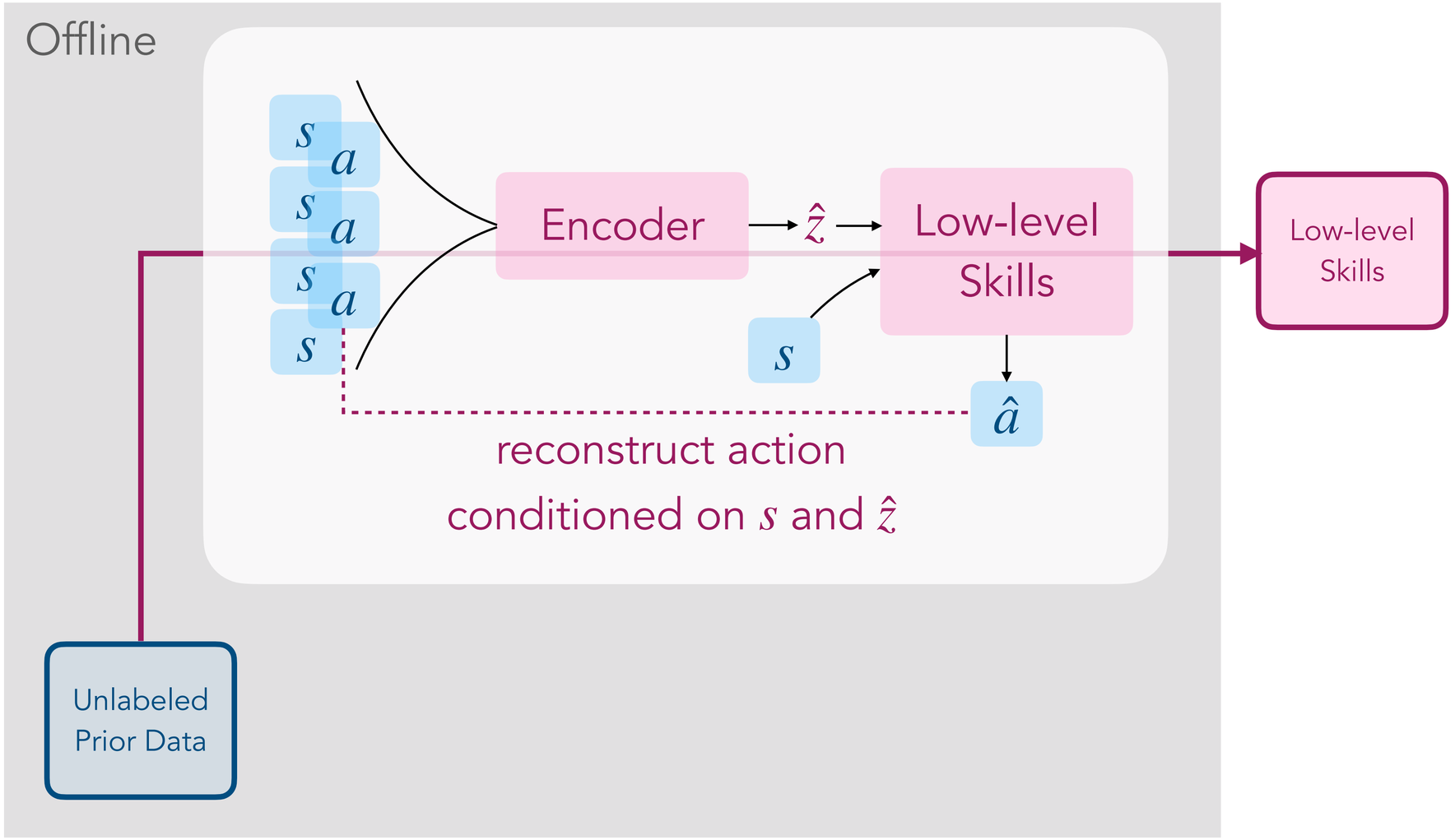
(1) Extract **low-level task-agnostic behaviors** offline with VAE skill pre-training to reduce task horizon



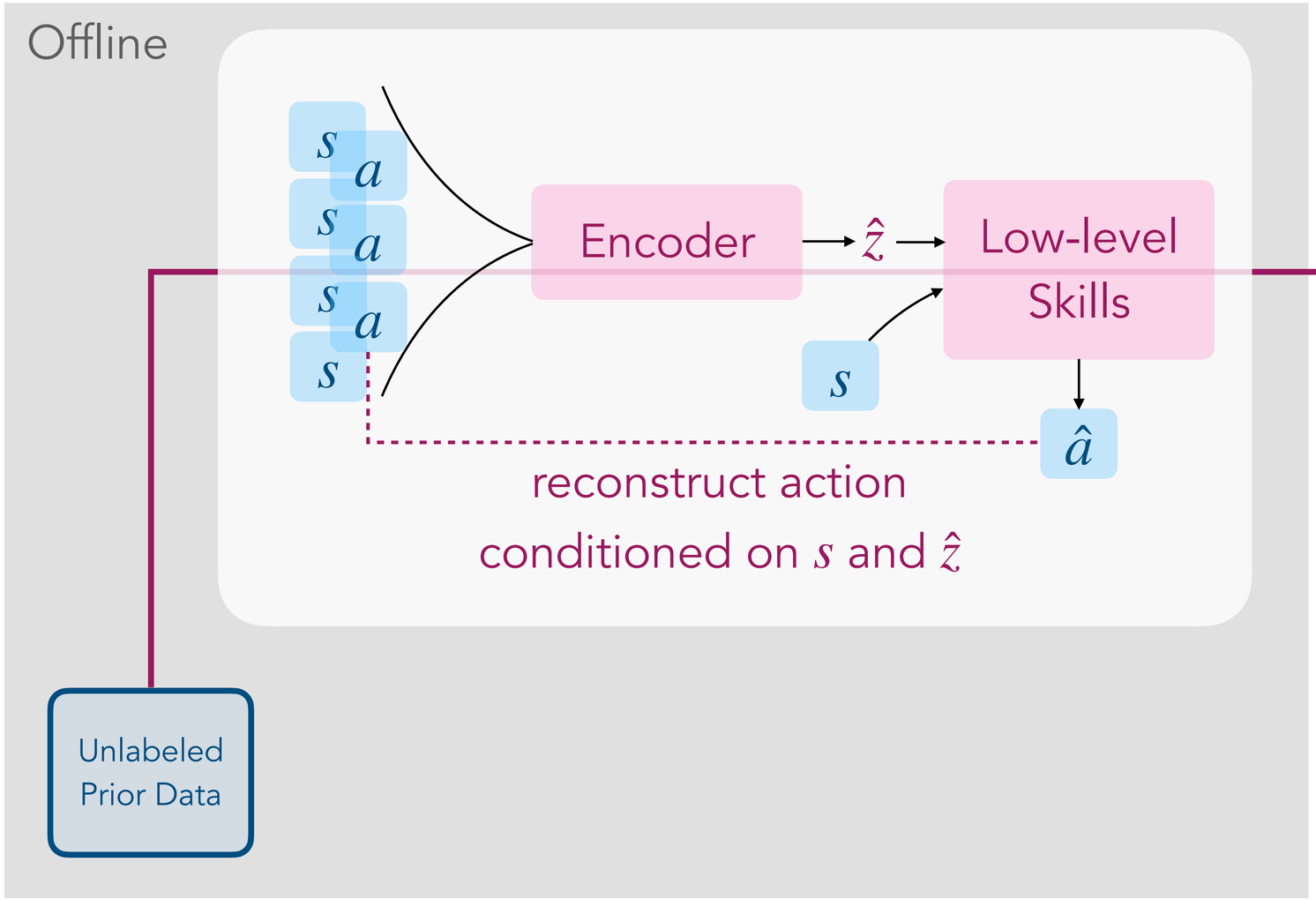


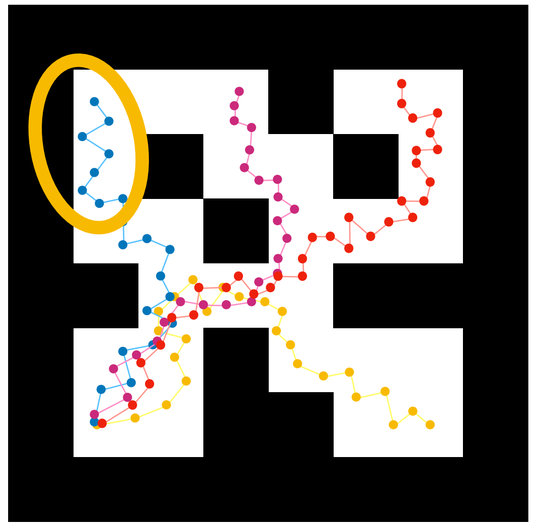
(1) Extract **low-level task-agnostic behaviors** offline
with VAE skill pre-training to reduce task horizon



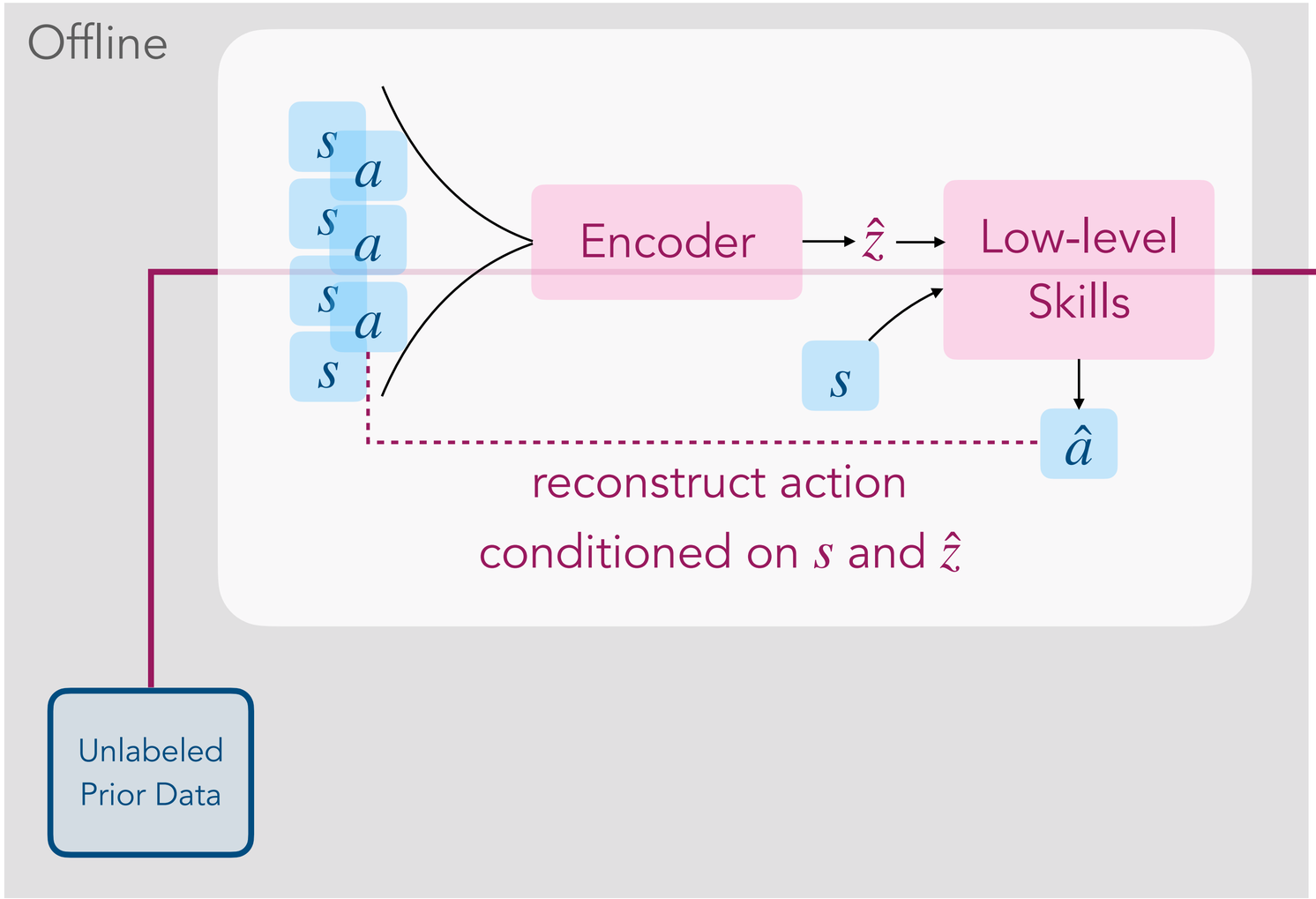


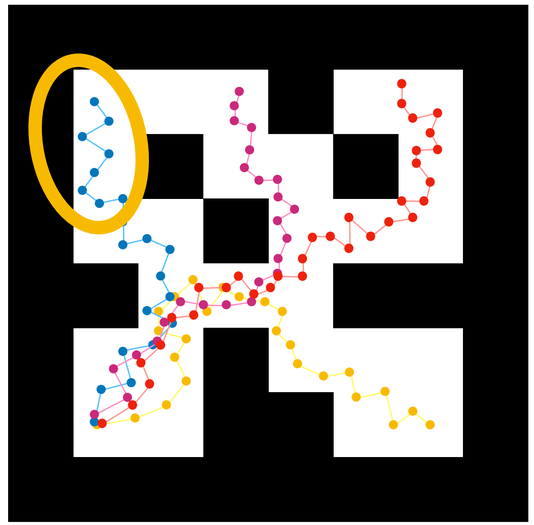
(1) Extract **low-level task-agnostic behaviors** offline
with VAE skill pre-training to reduce task horizon

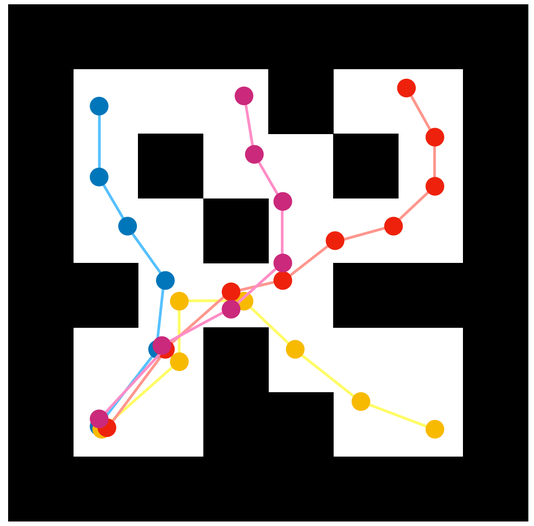




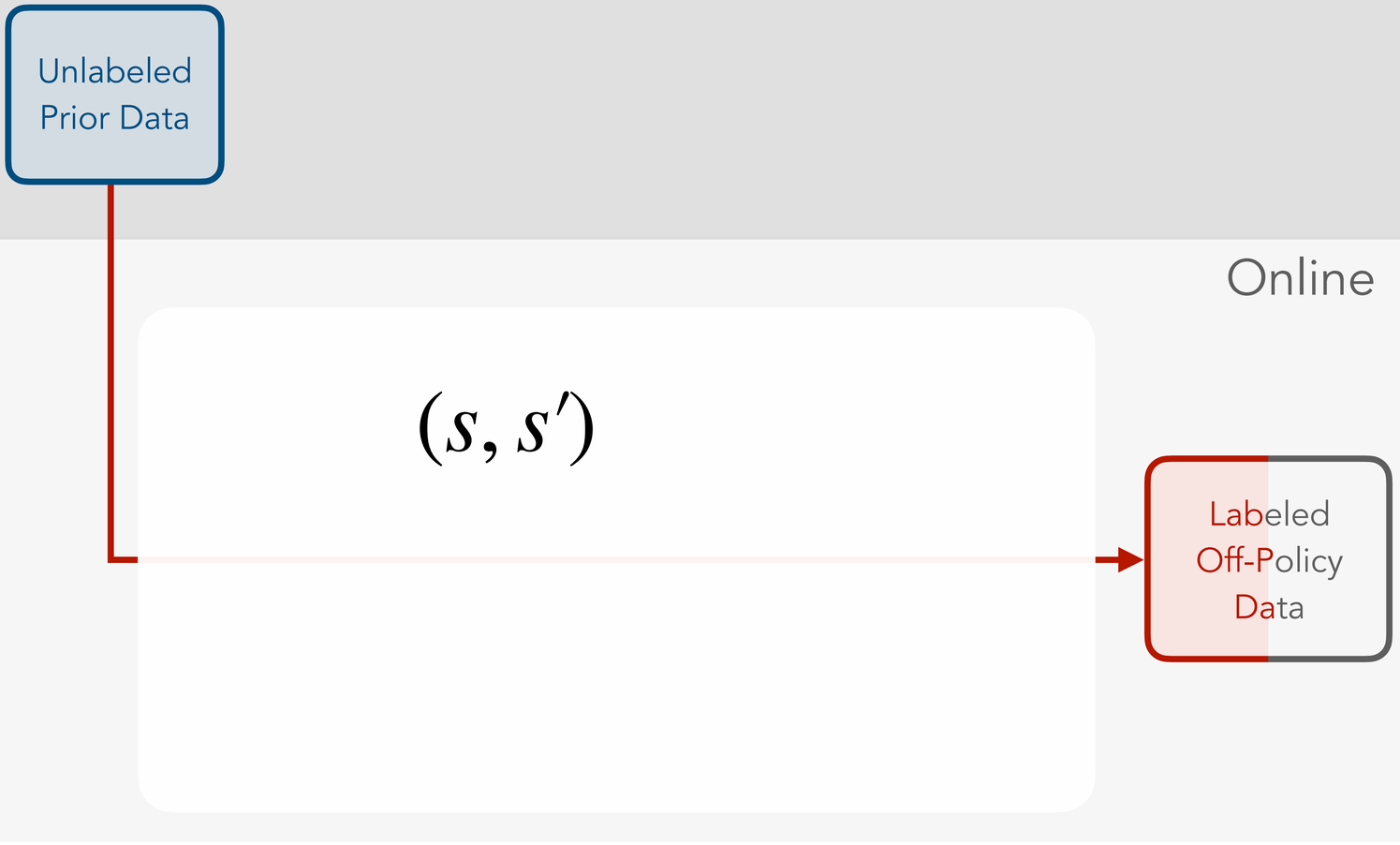
(1) Extract **low-level task-agnostic behaviors** offline
with VAE skill pre-training to reduce task horizon

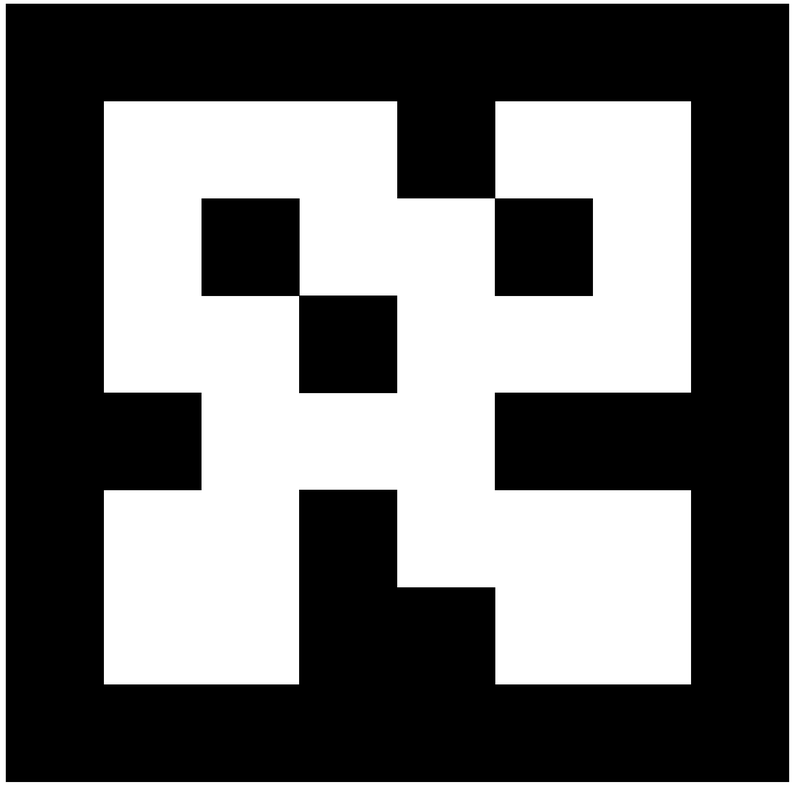






(2) Pseudo-label transitions with **optimistic rewards** to leverage the prior data as **additional off-policy data** and encourage online exploration





(2) Pseudo-label transitions with **optimistic rewards** to leverage the prior data as **additional off-policy data** and encourage online exploration

Unlabeled Prior Data

Online

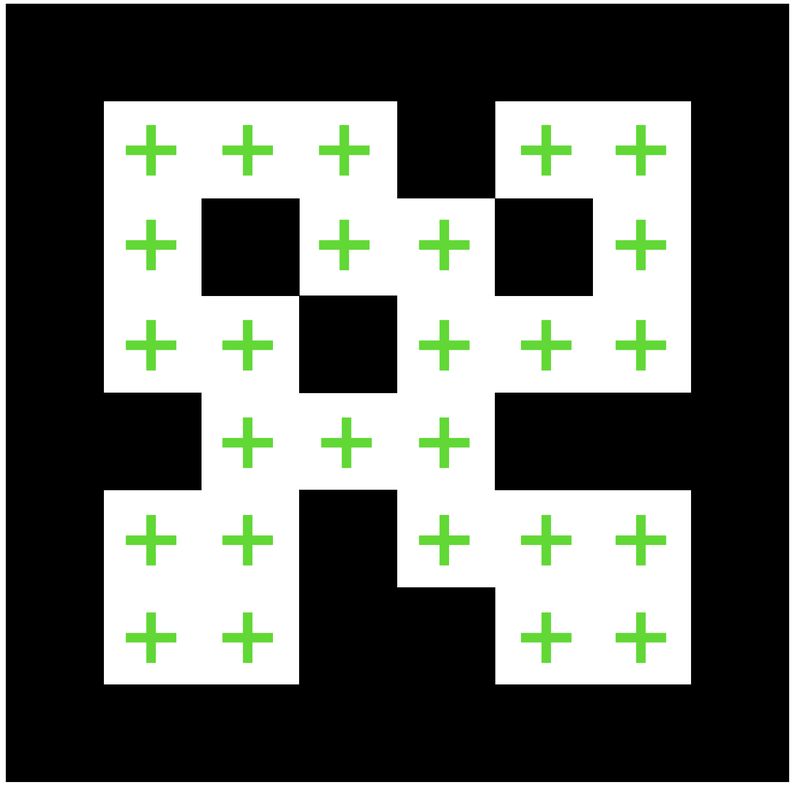
$$(s, \hat{z}, r^{+}(s, \hat{z}), s')$$

estimated skill latent

optimistic reward

(RND + online reward model)

Labeled Off-Policy Data



(2) Pseudo-label transitions with **optimistic rewards** to leverage the prior data as **additional off-policy data** and encourage online exploration

Unlabeled Prior Data

Online

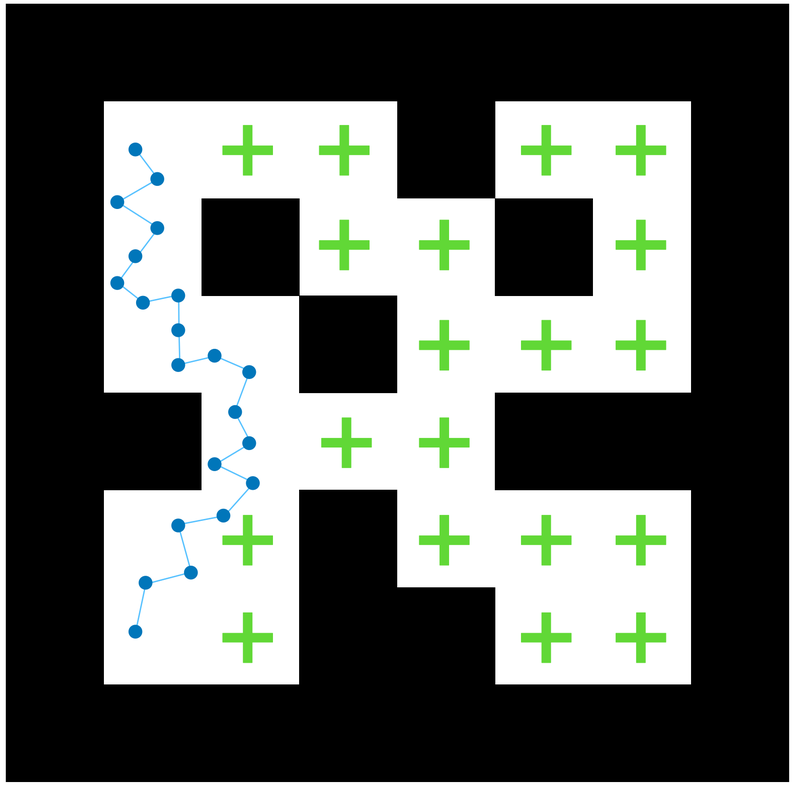
$$(s, \hat{z}, r^{+}(s, \hat{z}), s')$$

estimated skill latent

optimistic reward

(RND + online reward model)

Labeled Off-Policy Data



(2) Pseudo-label transitions with **optimistic rewards** to leverage the prior data as **additional off-policy data** and encourage online exploration

Unlabeled Prior Data

Online

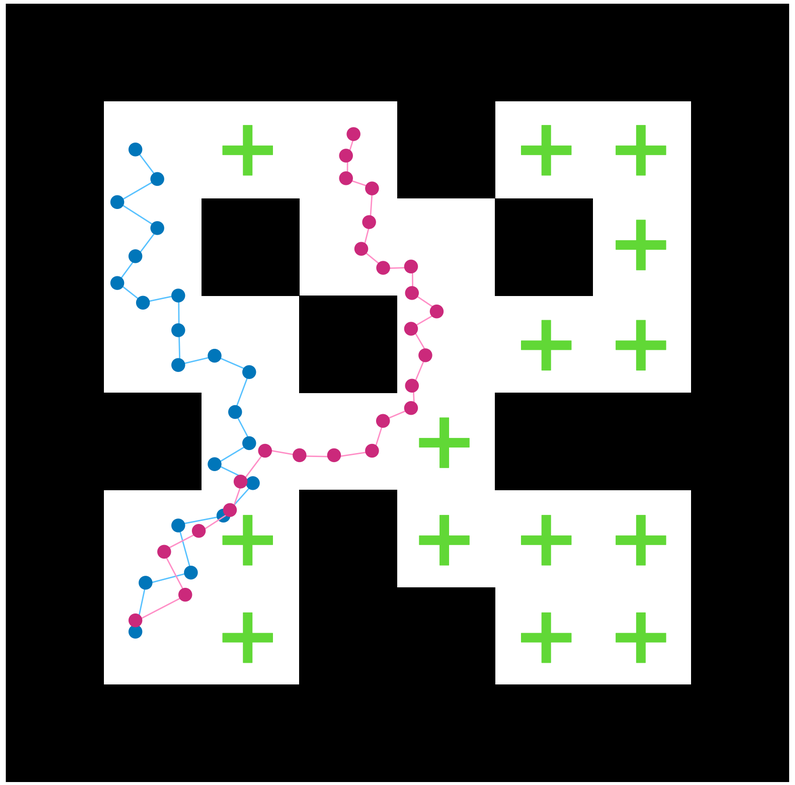
$$(s, \hat{z}, r^{+}(s, \hat{z}), s')$$

estimated skill latent

optimistic reward

(RND + online reward model)

Labeled Off-Policy Data



(2) Pseudo-label transitions with **optimistic rewards** to leverage the prior data as **additional off-policy data** and encourage online exploration

Unlabeled Prior Data

Online

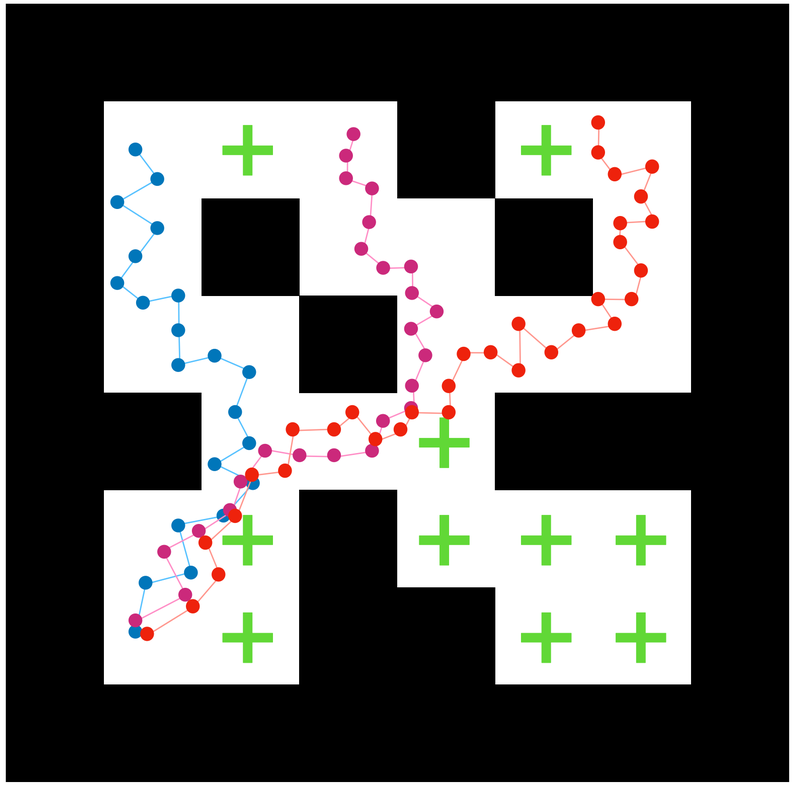
$$(s, \hat{z}, r^{+}(s, \hat{z}), s')$$

estimated skill latent

optimistic reward

(RND + online reward model)

Labeled Off-Policy Data



(2) Pseudo-label transitions with **optimistic rewards** to leverage the prior data as **additional off-policy data** and encourage online exploration

Unlabeled Prior Data

Online

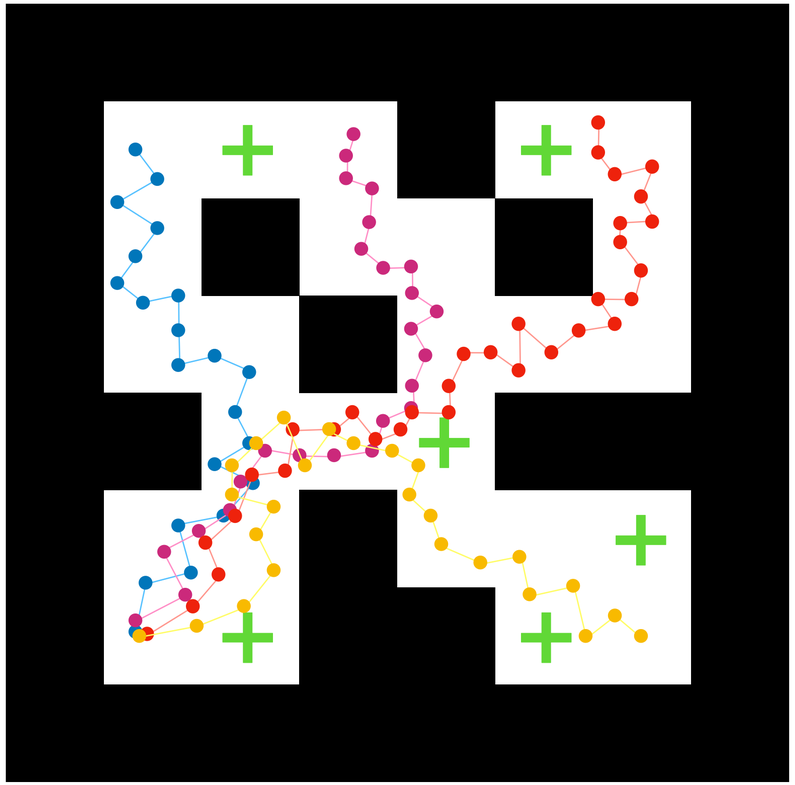
$$(s, \hat{z}, r^{+}(s, \hat{z}), s')$$

estimated skill latent

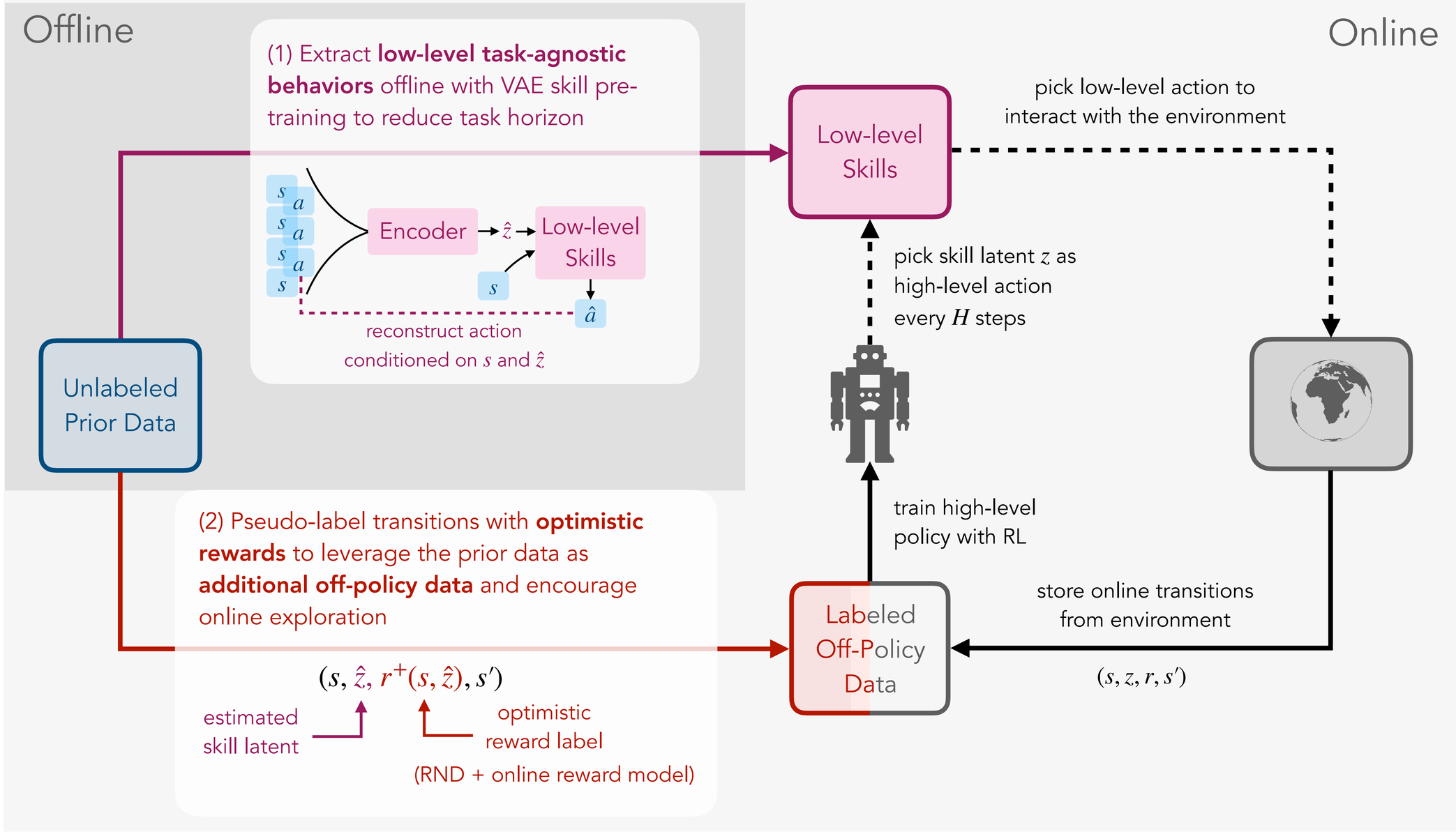
optimistic reward

(RND + online reward model)

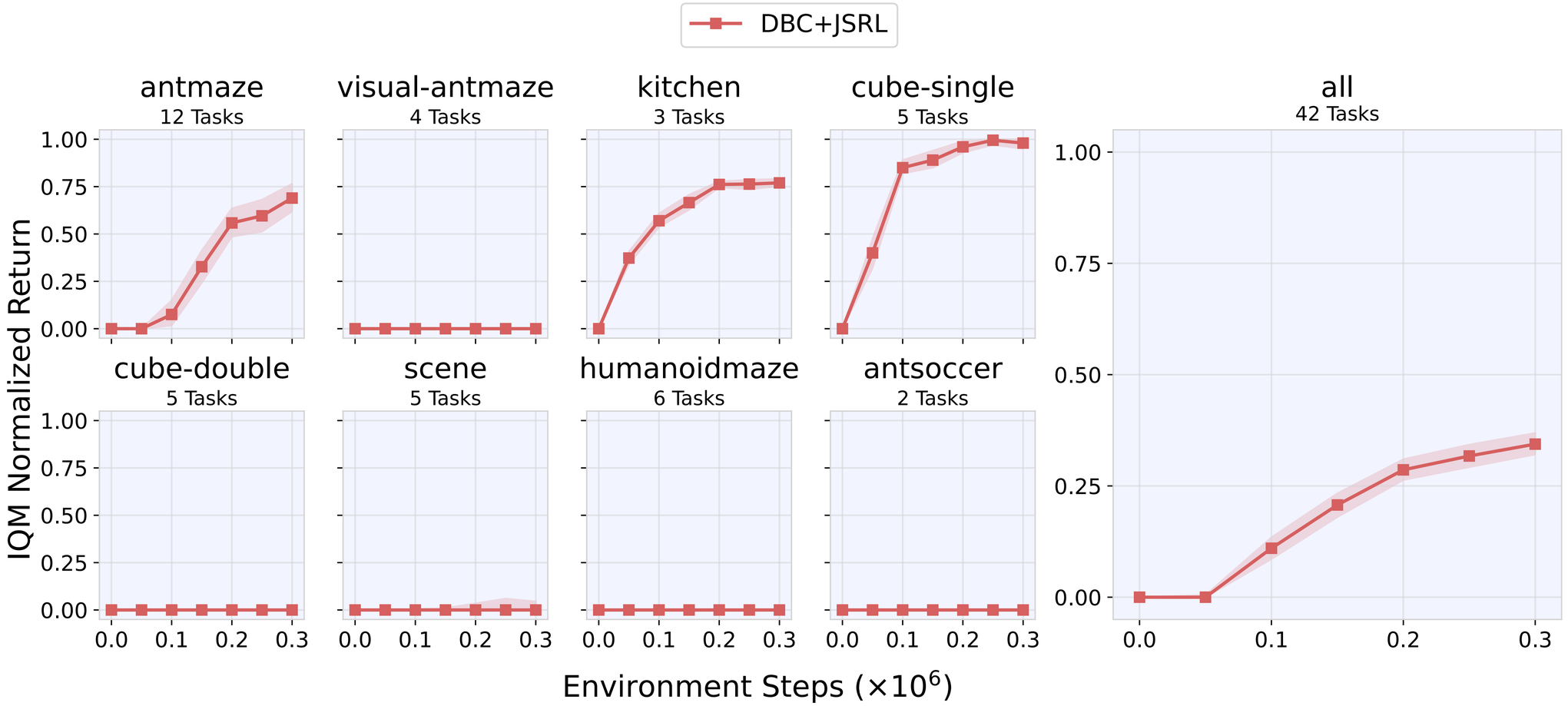
Labeled Off-Policy Data



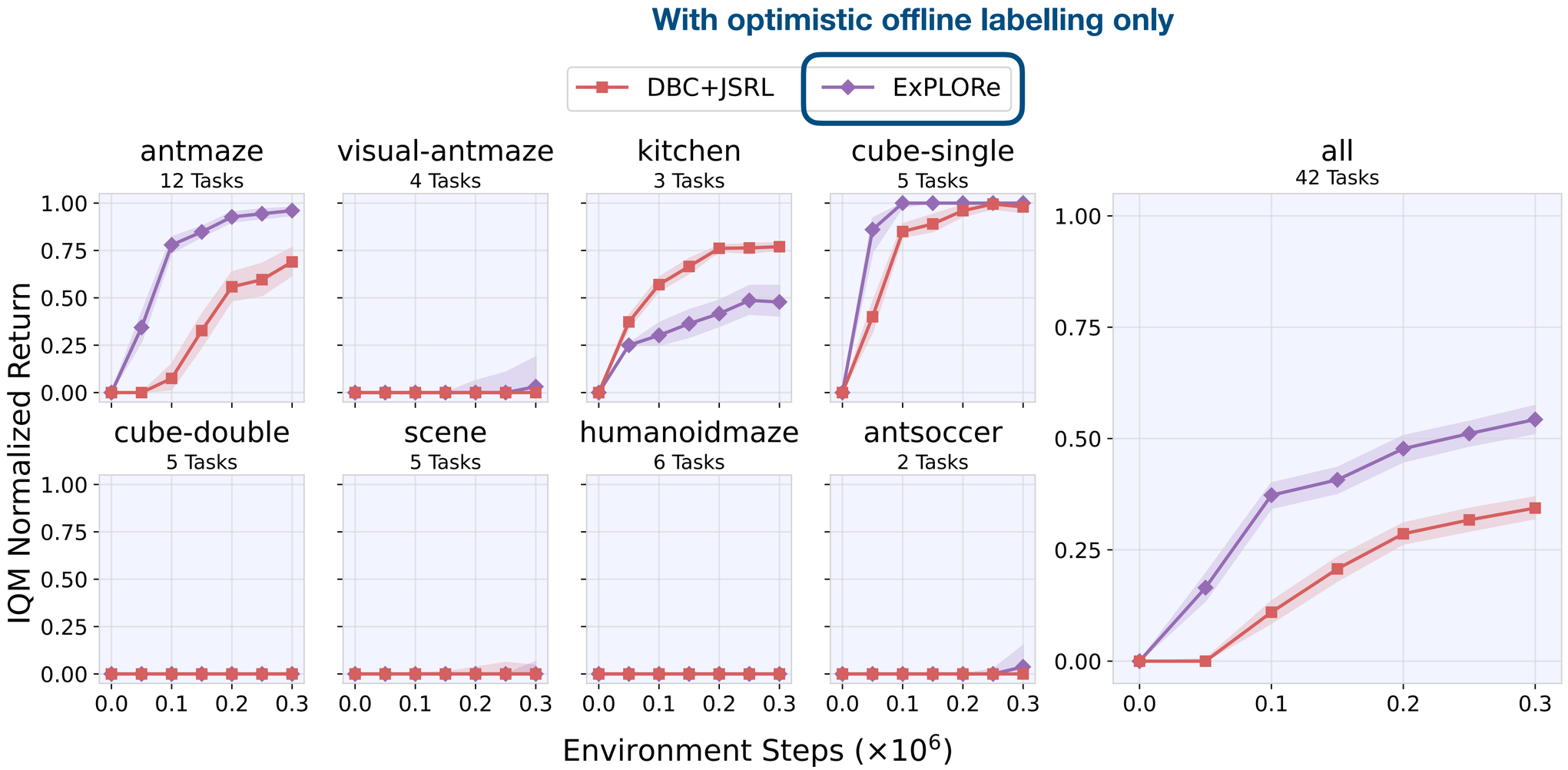
SUPE: **S**kills from **U**nlabeled **P**rior Data for **E**xploration



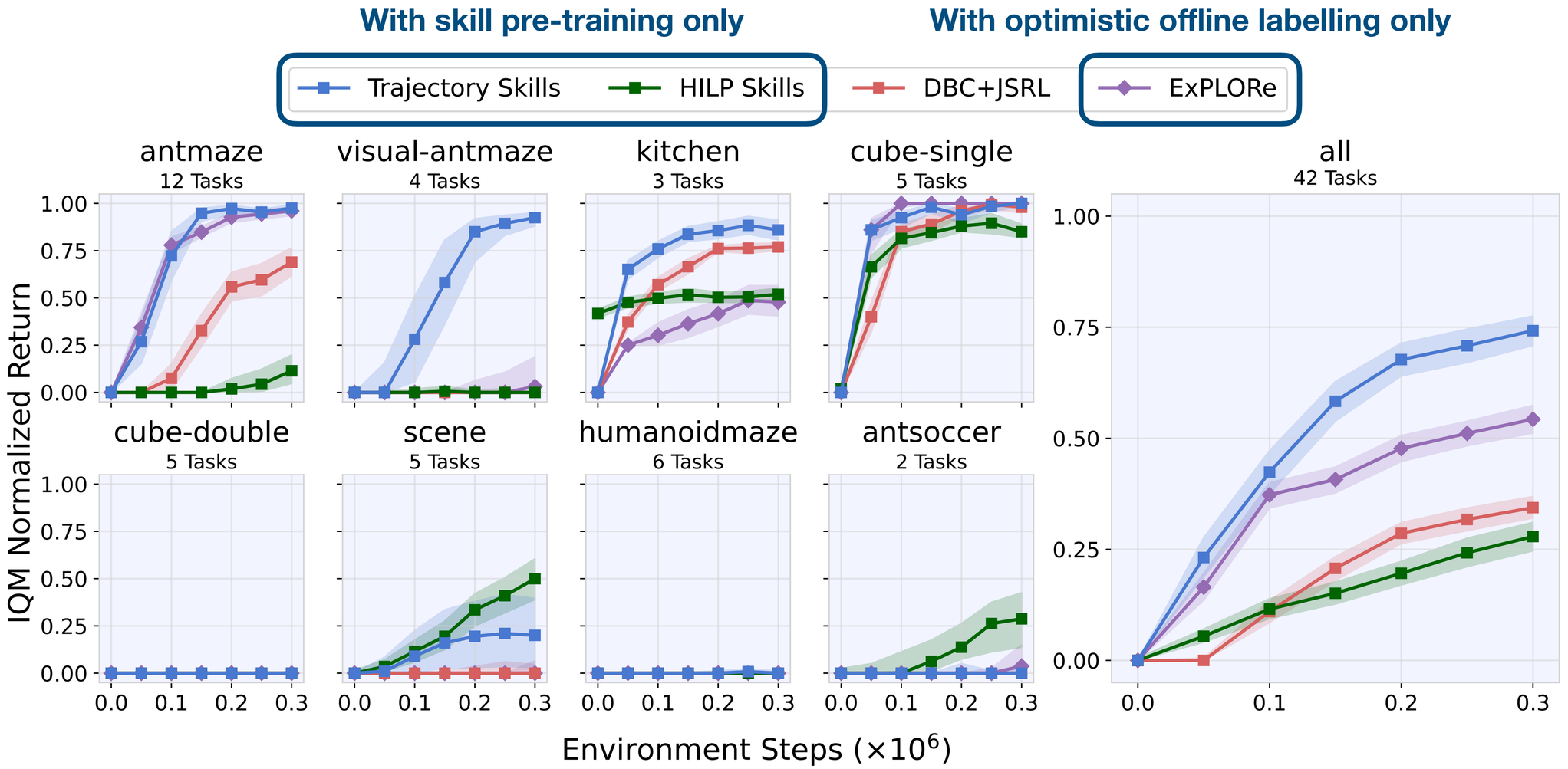
How well does SUPE do?



How well does SUPE do?



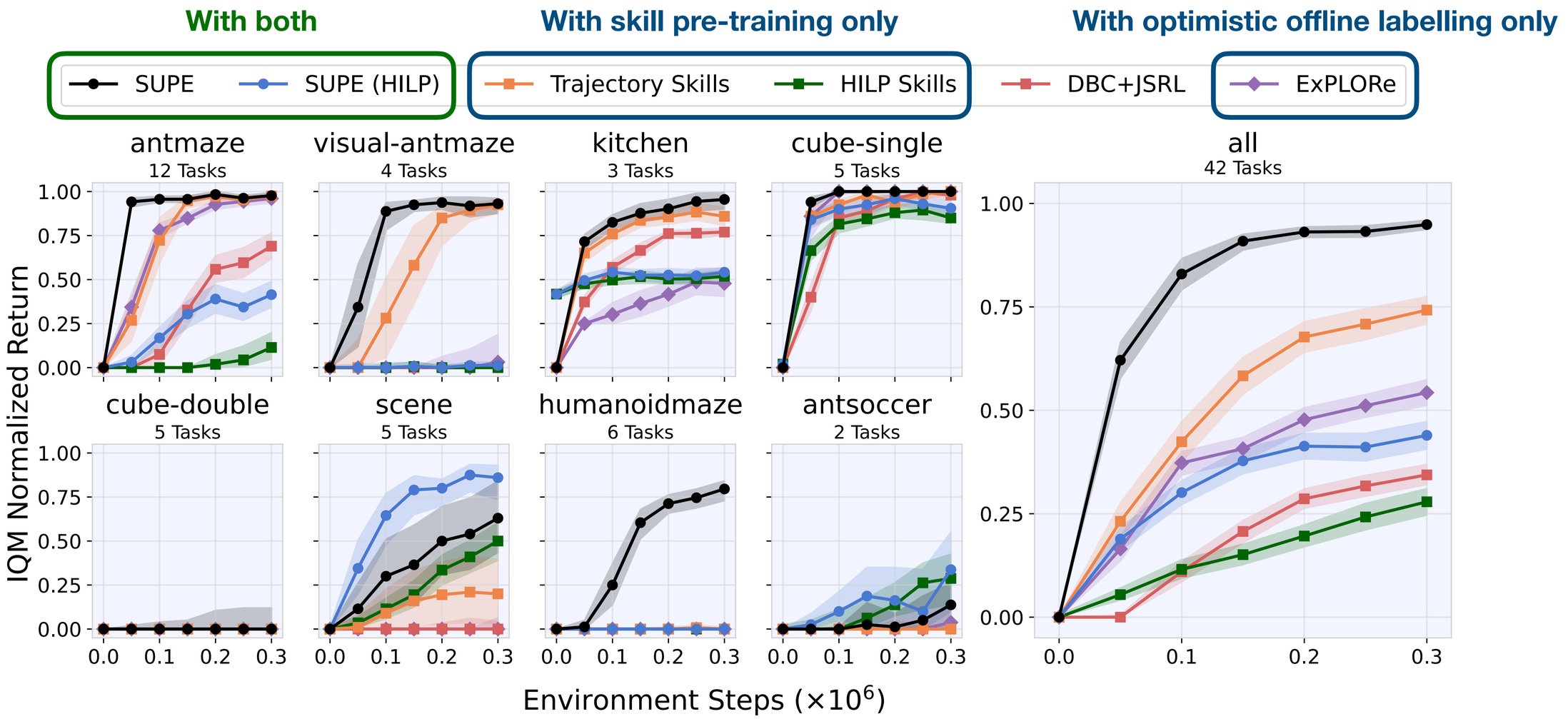
How well does SUPE do?



Anurag Ajay, Aviral Kumar, Pulkit Agrawal, Sergey Levine, and Ofir Nachum. OPAL: Offline primitive discovery for accelerating offline reinforcement learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum? id=V69LGwJ0lIN.

Seohong Park, Tobias Kreiman, and Sergey Levine. Foundation policies with hilbert representations. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/ forum?id=LhNsSaAKub.

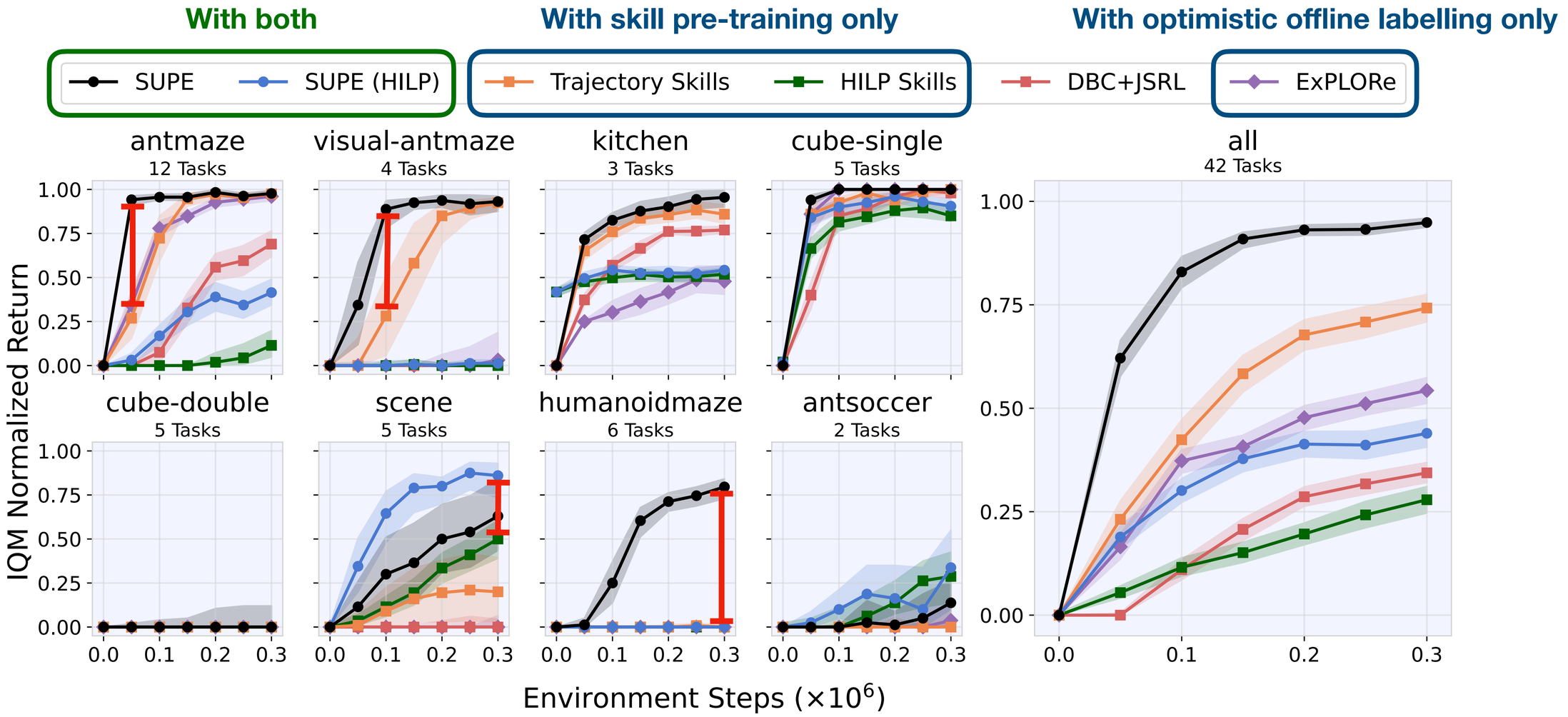
How well does *SUPE* do?



Anurag Ajay, Aviral Kumar, Pulkit Agrawal, Sergey Levine, and Ofir Nachum. OPAL: Offline primitive discovery for accelerating offline reinforcement learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum? id=V69LGwJ0lIN.

Seohong Park, Tobias Kreiman, and Sergey Levine. Foundation policies with hilbert representations. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/ forum?id=LhNsSaAKub.

How well does ***SUPE*** do?



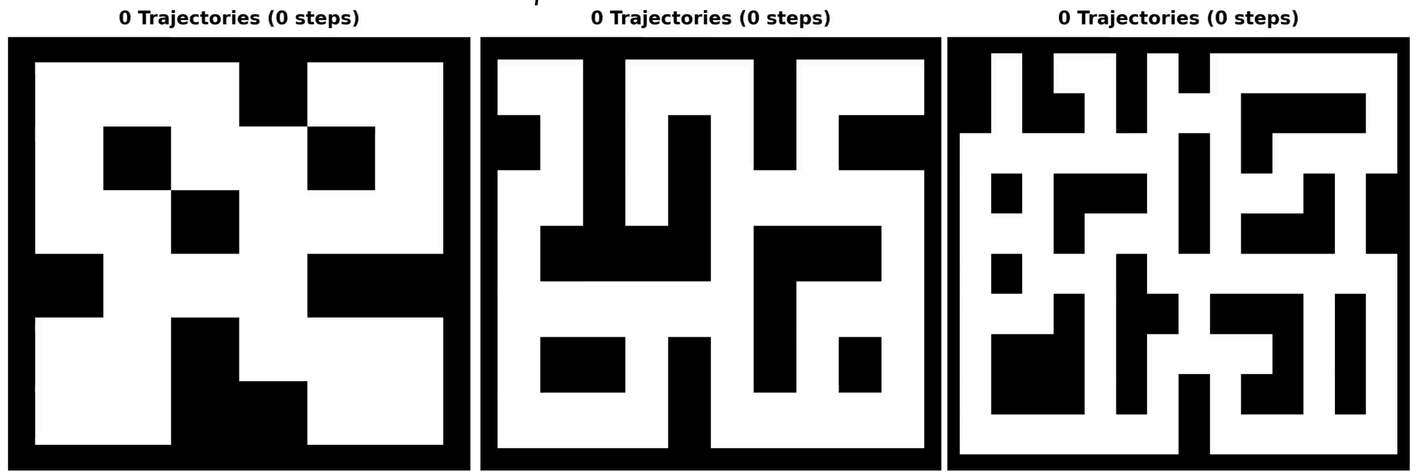
Anurag Ajay, Aviral Kumar, Pulkit Agrawal, Sergey Levine, and Ofir Nachum. OPAL: Offline primitive discovery for accelerating offline reinforcement learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum? id=V69LGwJ0lIN.

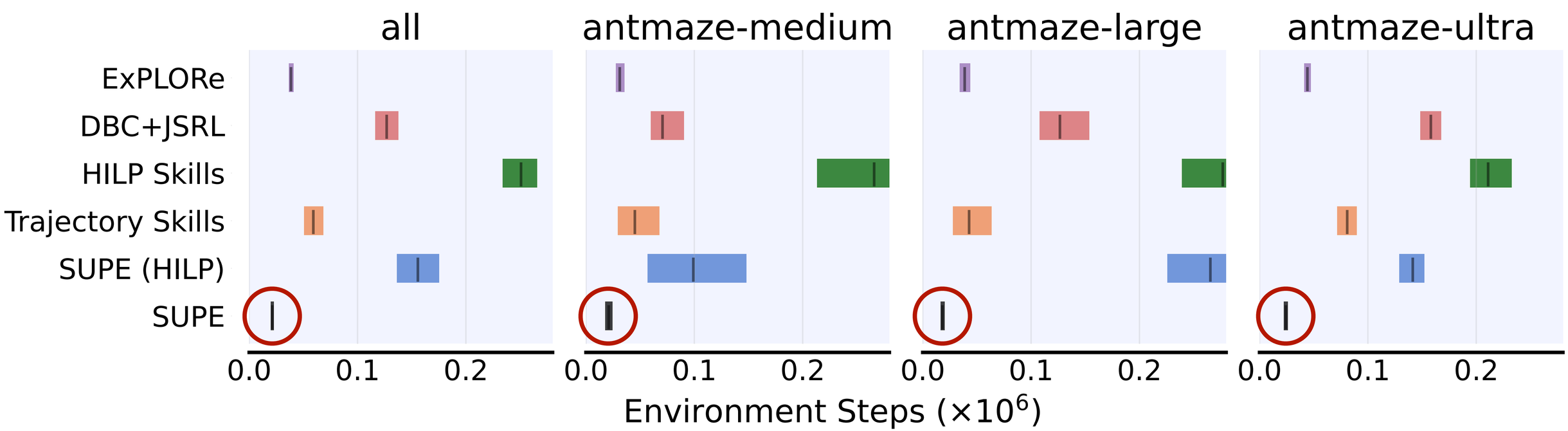
Seohong Park, Tobias Kreiman, and Sergey Levine. Foundation policies with hilbert representations. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/ forum?id=LhNsSaAKub.

Does *SUPE* *actually* explore better?

Metric: # of steps it takes for the agent to reaches the goal for the first time

Exploration Visualization

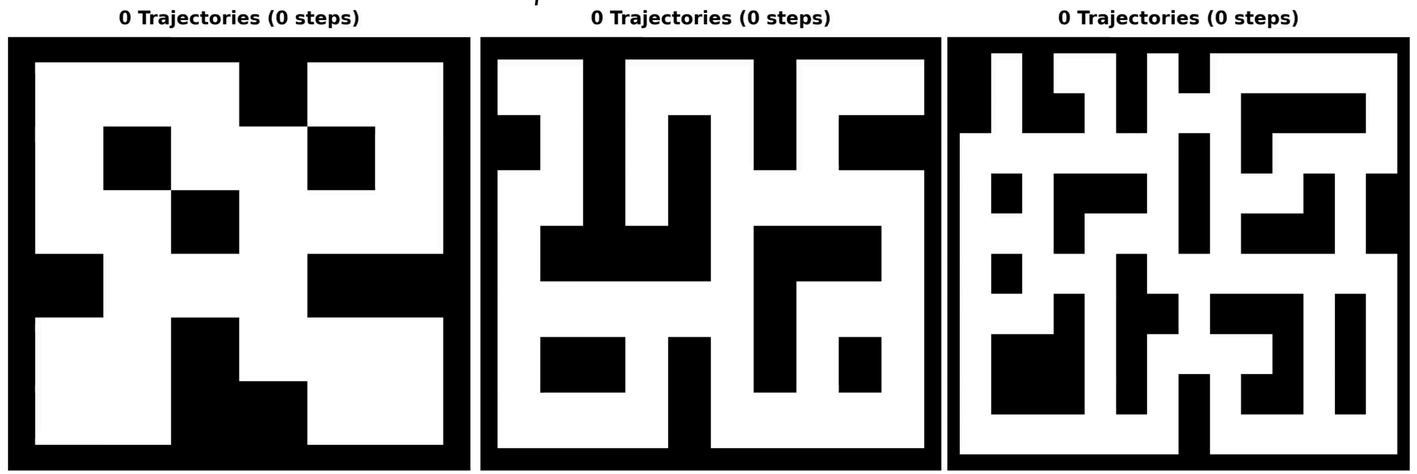


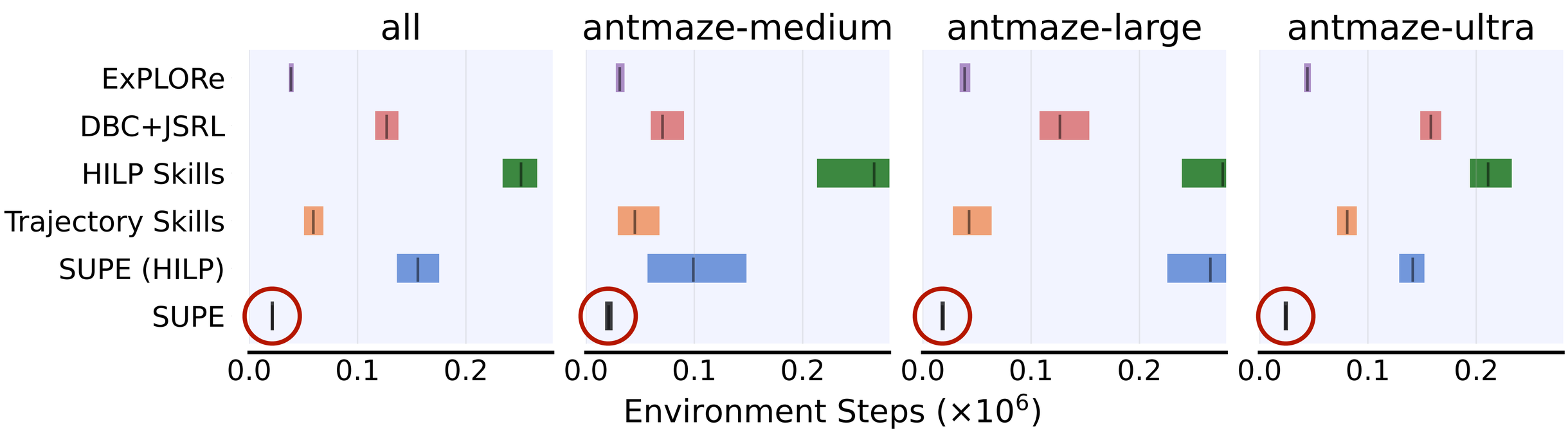


Does *SUPE* *actually* explore better?

Metric: # of steps it takes for the agent to reaches the goal for the first time

Exploration Visualization





Thank you for your time!!

We have open-sourced our code, and it is available at https://github.com/rail-berkeley/SUPE.