IR-VIC: Unsupervised Discovery of Sub-goals for Transfer in RL

Nirbhay Modhe[1] Prithvijit Chattopadhyay[1] Mohit Sharma[1] Abhishek Das[1]

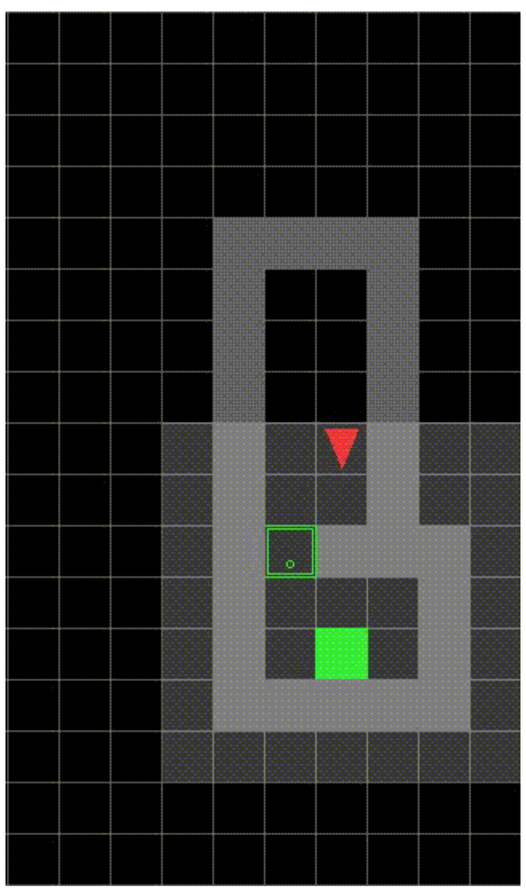



Devi Parikh[1,2] Dhruv Batra[1,2] Ramakrishna Vedantam[2]

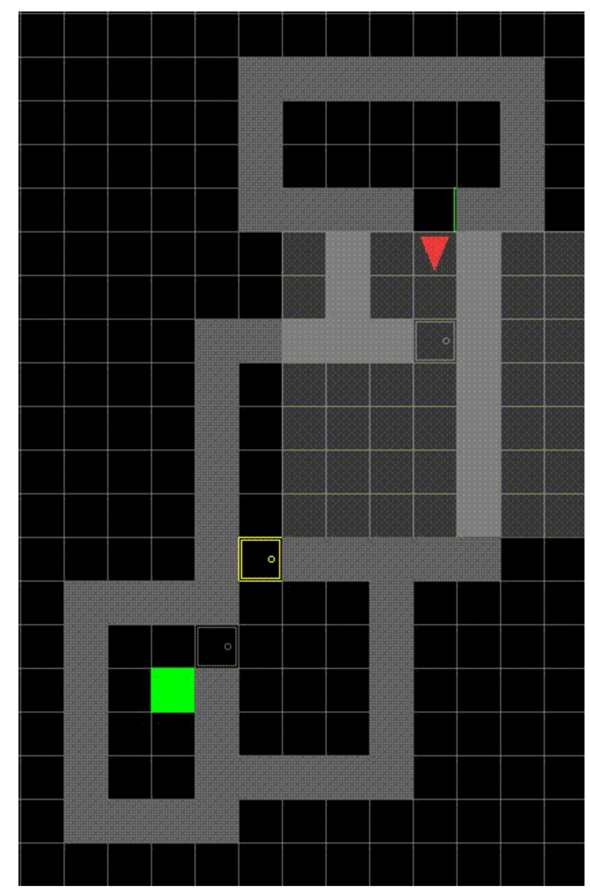


[1]Georgia Institute of Technology
[2]Facebook AI Research

Exploration with Sparse Rewards



Sparse reward, +1
for reaching goal



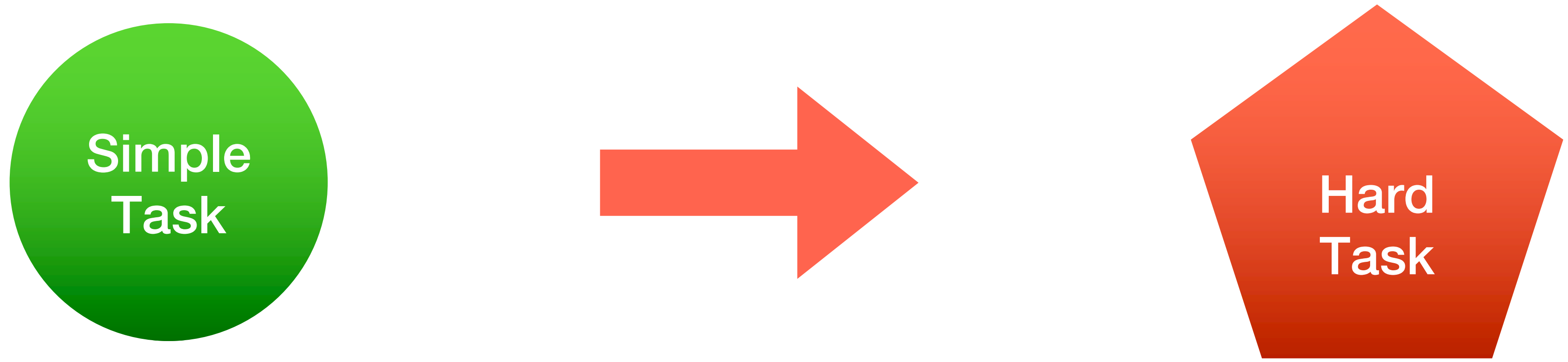
Gif Source: “Minimalistic Gridworld Environment for OpenAI Gym”, 2018, github.com/maximecb/gym-minigrid

Transfer of Sub-goals

"...decomposing a complex problem into a set of simpler ones"
- McGovern & Barto, 2001[1]

[1] McGovern, Amy, and Andrew G. Barto. "Automatic discovery of subgoals in reinforcement learning using diverse density." ICML 2001

Transfer of Sub-goals

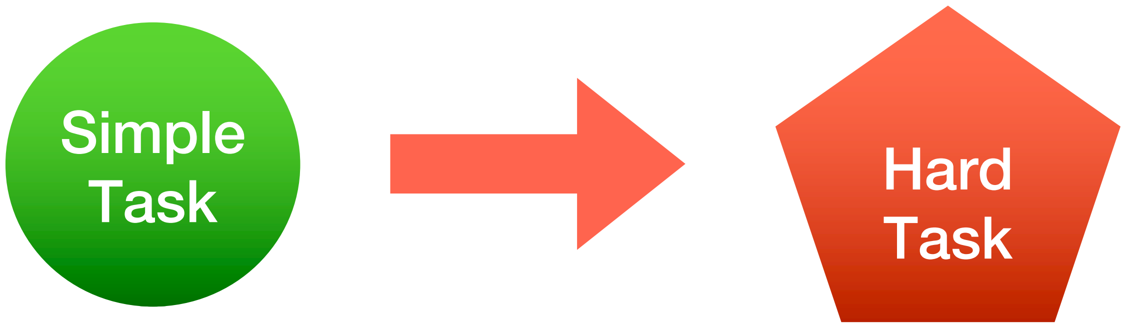


Learn sub-goal detector in source environment

Transfer sub-goal detector to target environment

Images are in CC0 Public Domain, source: openclipart.org

Challenges of Transfer



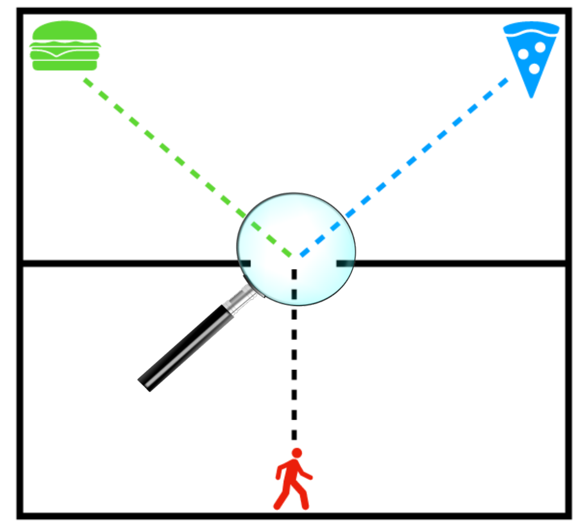
Challenges for sub-goal transfer:

- Detector should be **easy to train** in source environments
- Transferable to **novel target environments**
- **Improve performance** in target environment's task

Information-based Sub-goals

Peaks in **Relevant Goal Information (RGI)**

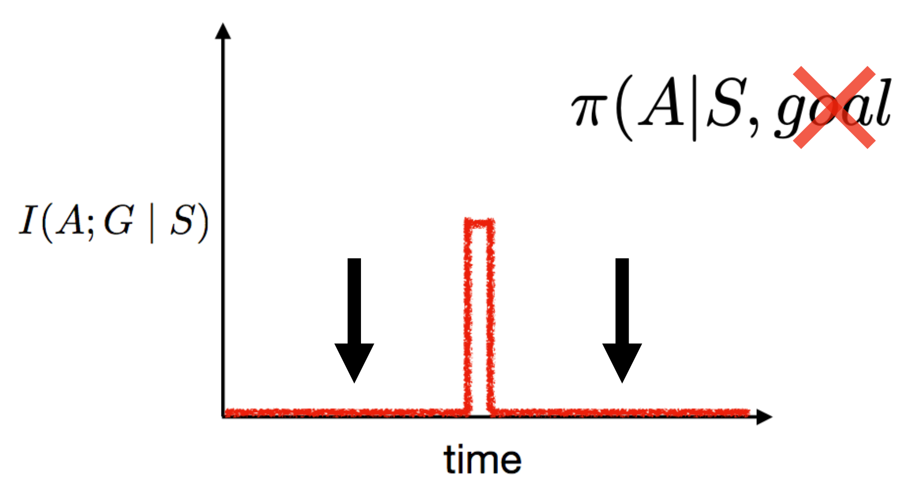
Goal-informed actions



G ~ (,)

Default Behavior (**Low RGI**)

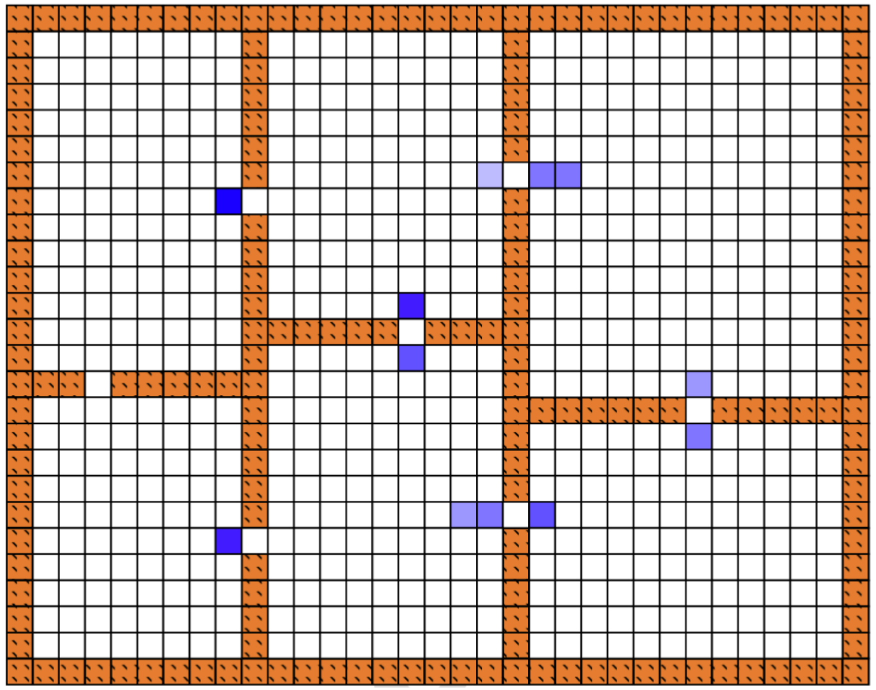
Goal-independent actions



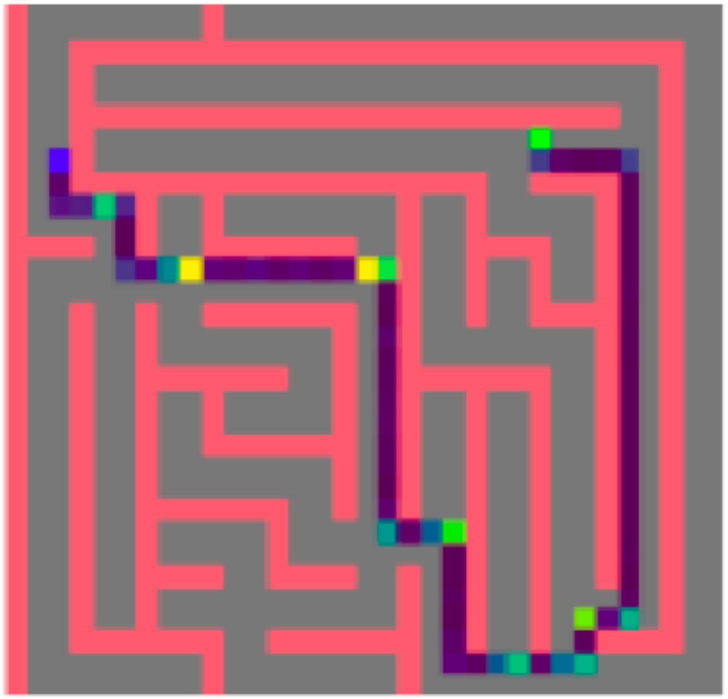
Figures adapted from: DJ Strouse, "Hierarchical Reinforcement Learning via Information Bottleneck"

Sub-goal Identification

Sub-goals discovered with **goal-driven (extrinsic) rewards**



Dijk & Polani "*Grounding Subgoals in Information Transitions.*" 2011

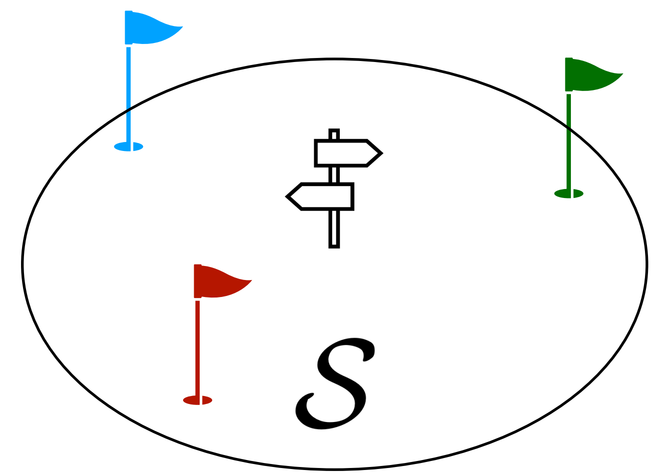


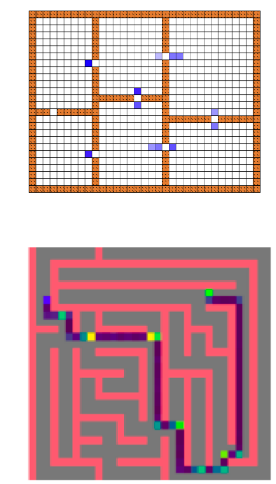
Goyal et. al. "*InfoBot: Transfer and Exploration via the Information Bottleneck.*" 2019

Sub-goal Identification

- Sub-goals discovered with **goal-driven (extrinsic) rewards**

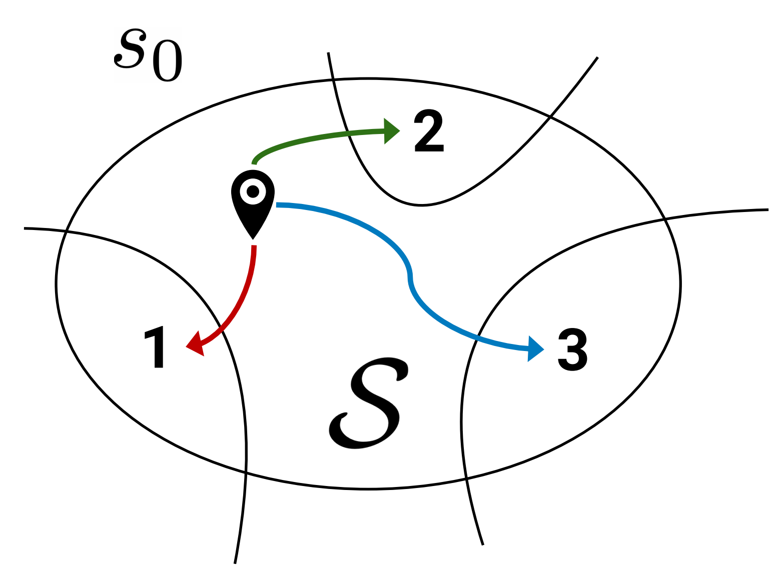
$$= f(\ \text{task}\ ,\ \text{environment}\ )$$





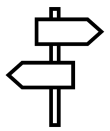
- Sub-goals discovered with **intrinsic rewards**

$$= f(\ \text{environment}\ )$$

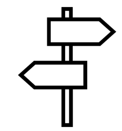


self-supervised goals

Sub-goal Identification

 $= f(\text{ task },\ \text{environment })$

- Require **task specification** in source environment
- May not generalize to **different tasks** in similar environments

 $= f(\text{ environment })$

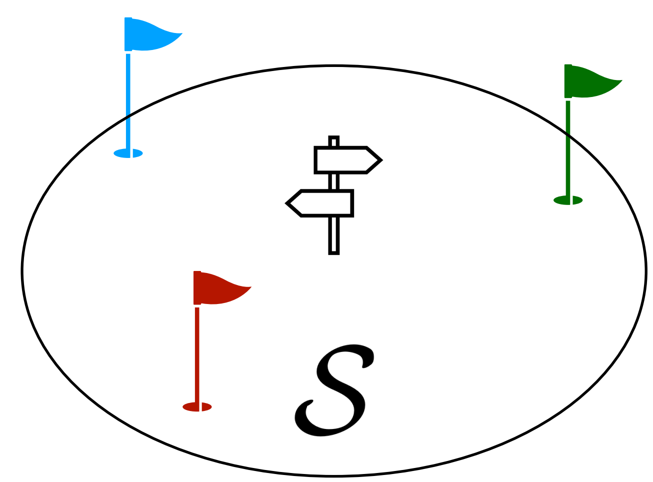
- Unsupervised, **task-independent** objective
- Exploits **environment structure** alone, generalizing better to similar environments

Unsupervised Sub-goal Discovery

Supervised goal: ↑ External Reward ↓ $I(\text{actions}; \text{goals})$

$$\max_{\pi_\theta} [\ r - \beta I(A; G)\ ]$$

Maximize task reward with penalty for using goal information



Unsupervised Sub-goal Discovery

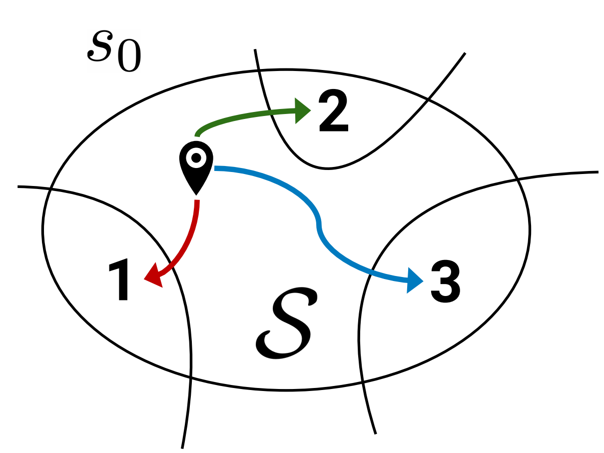
IR-VIC: Sub-goal discovery **without external tasks**

Supervised goal: External Reward $I(\text{actions}; \text{goals})$

Unsupervised goal: Intrinsic Control $I(\text{actions}; \text{options})$

Learn intrinsic options

Look at option sparingly



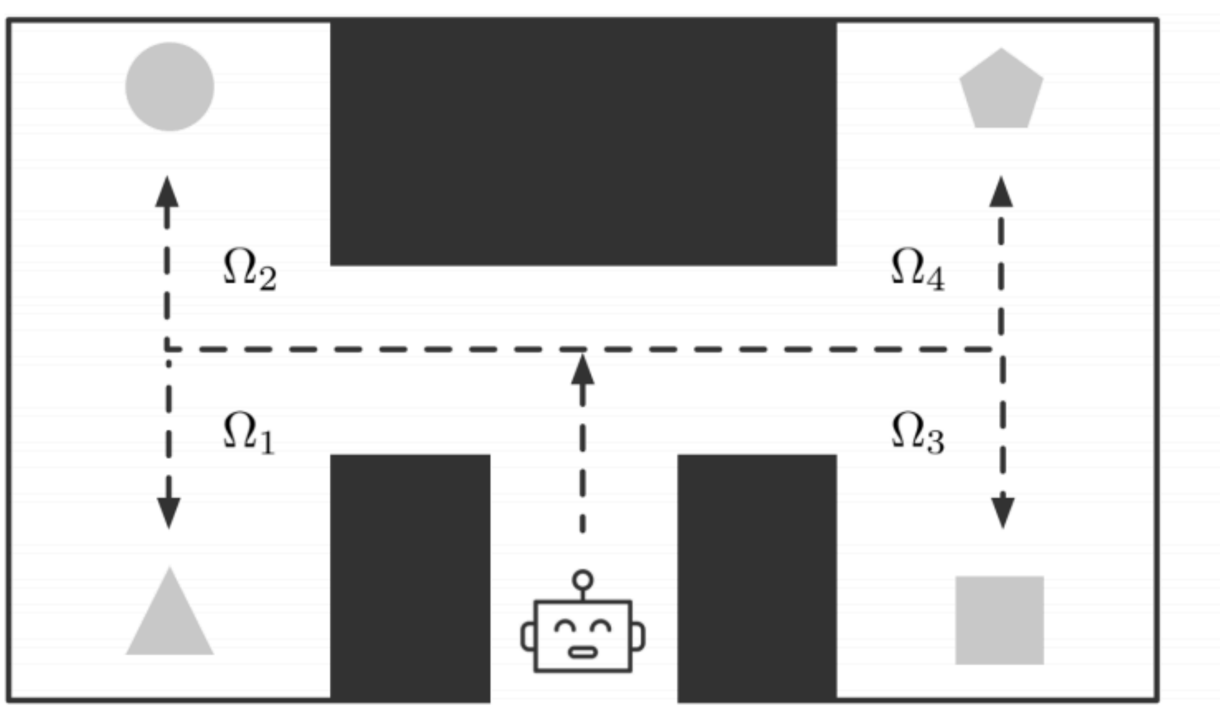
Maximize intrinsic reward with penalty for using option information

Learning Intrinsic Options

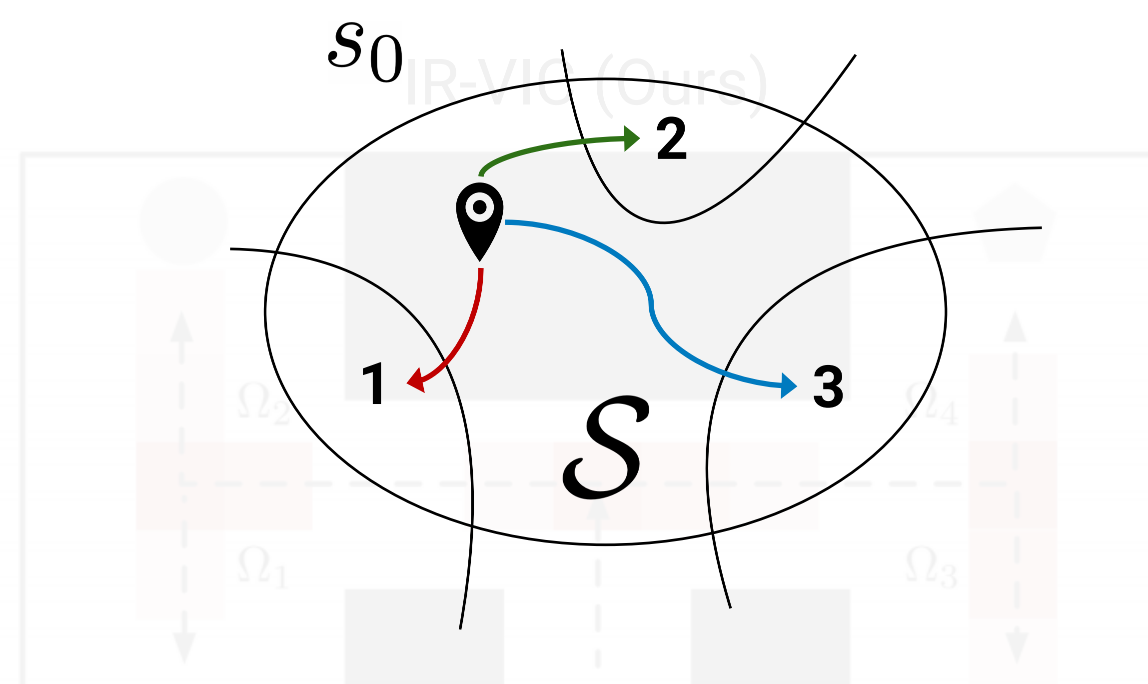
Unsupervised goal: ↑ Intrinsic Control ↓ $I(\text{actions}; \text{options})$

Learn intrinsic options

VIC: Variational Intrinsic Control[1]
(Gregor et al., 2016)



Ω : option



Learn a set of options terminating at distinct final states

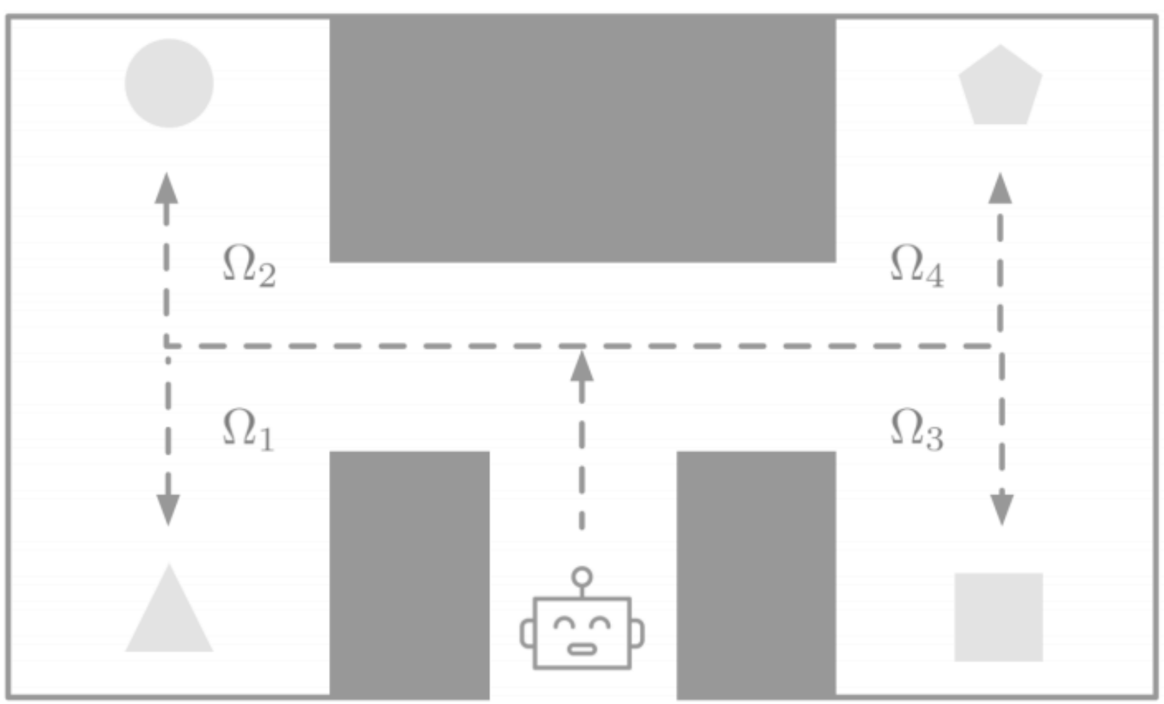
[1] Gregor, Karol, Danilo Jimenez Rezende, and Daan Wierstra. "Variational intrinsic control." arXiv:1611.07507 (2016).

Information Minimization

Unsupervised goal: ↑ Intrinsic Control ↓ $\boxed{I(\text{actions}; \text{options})}$

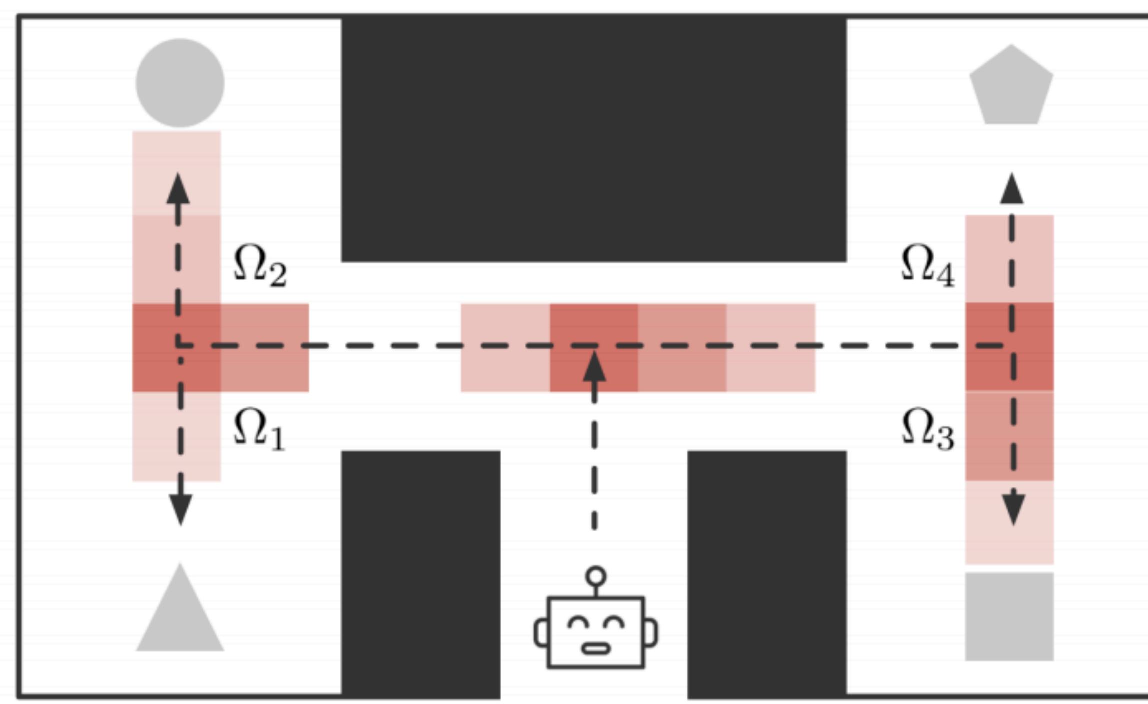
Enforce "default behavior" / option-independent actions

VIC: Variational Intrinsic Control[1]
(Gregor et al., 2016)



Ω : option

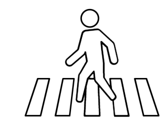
IR-VIC (Ours)



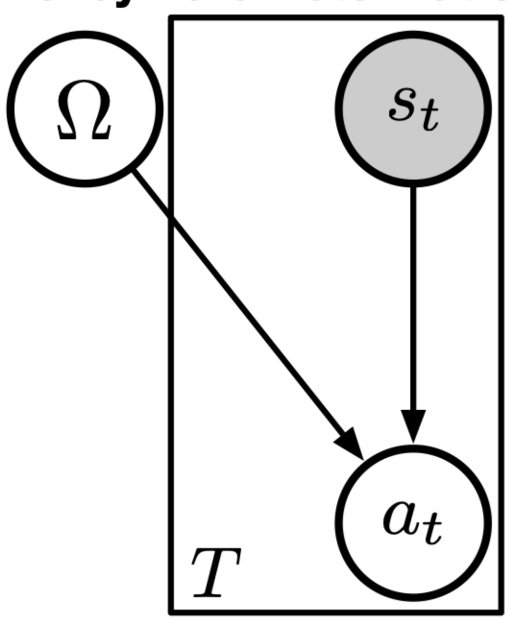
[1] Gregor, Karol, Danilo Jimenez Rezende, and Daan Wierstra. "Variational intrinsic control." arXiv:1611.07507 (2016).

Information Minimization

Unsupervised goal: Intrinsic Control $I(\text{actions}; \text{options})$

Enforce "default behavior" / option-independent actions 

Policy Parameterization



policy

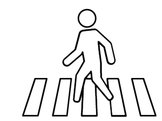
$\pi(a_t|s_t, \Omega)$

$$I(a_t; \Omega|s_t) \leq I(z_t; \Omega|s_t)$$

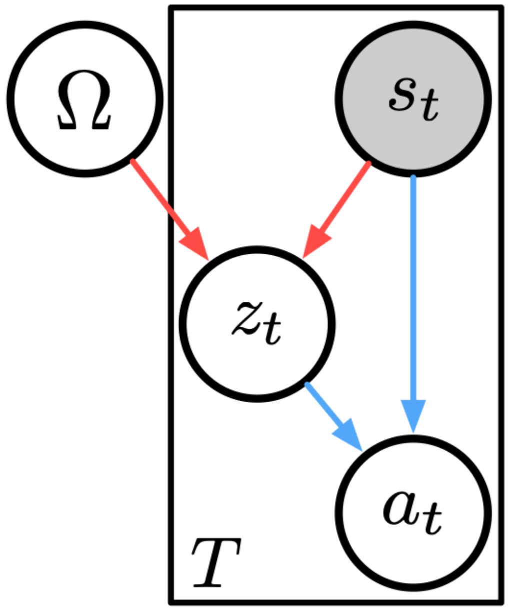
Ideal: Minimize action-goal information

Information Minimization

Unsupervised goal: Intrinsic Control $I(\text{actions}; \text{options})$

Enforce "default behavior" / option-independent actions 

Policy Parameterization



policy
$\pi(a_t|s_t, \Omega)$

encoder
$p_{enc}(z_t|s_t, \Omega)$

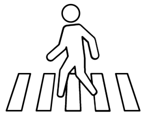
decoder
$p_{dec}(a_t|s_t, z_t)$

$$I(a_t; \Omega|s_t) \leq I(z_t; \Omega|s_t)$$

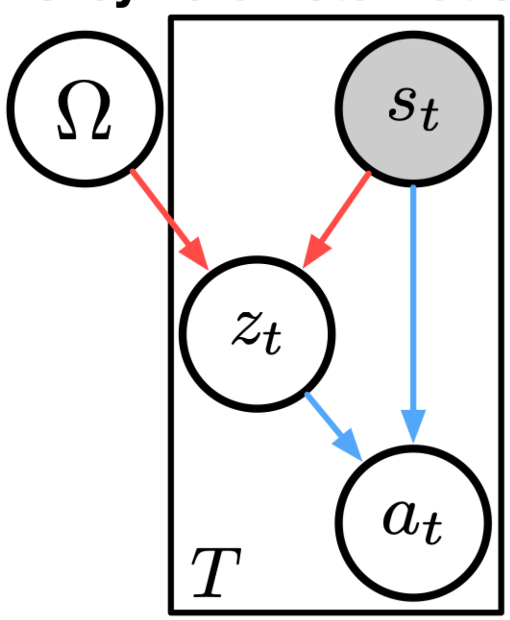
Practical: Minimize upper bound

Information Minimization

Unsupervised goal: Intrinsic Control $\boxed{I(\text{actions}; \text{options})}$

Enforce "default behavior" / option-independent actions 

Policy Parameterization



policy
$\pi(a_t|s_t, \Omega)$

encoder
$p_{enc}(z_t|s_t, \Omega)$

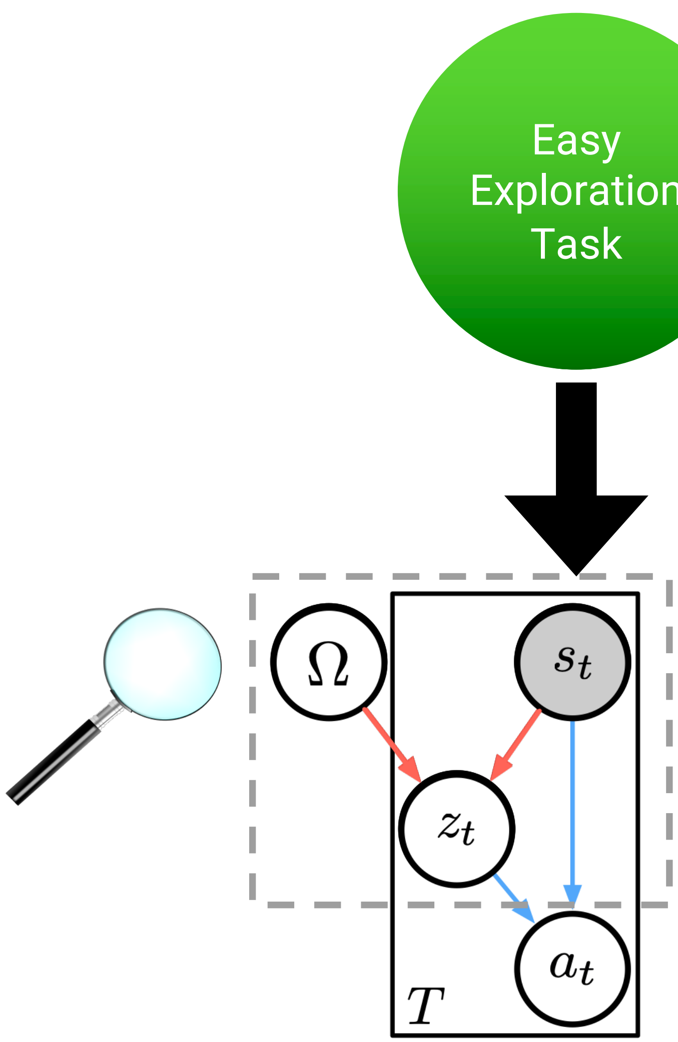
decoder
$p_{dec}(a_t|s_t, z_t)$

$$I(a_t; \Omega|s_t) \leq \boxed{I(z_t; \Omega|s_t)}$$

Sub-goals are peaks in MI 

Sub-goal Transfer

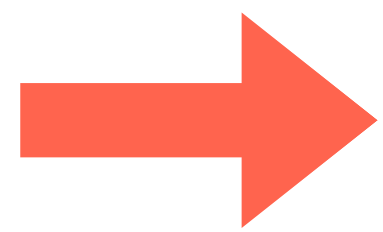
Learn sub-goal detector
for simple task



policy
$\pi(a_t|s_t, \Omega)$

encoder
$p_{enc}(z_t|s_t, \Omega)$

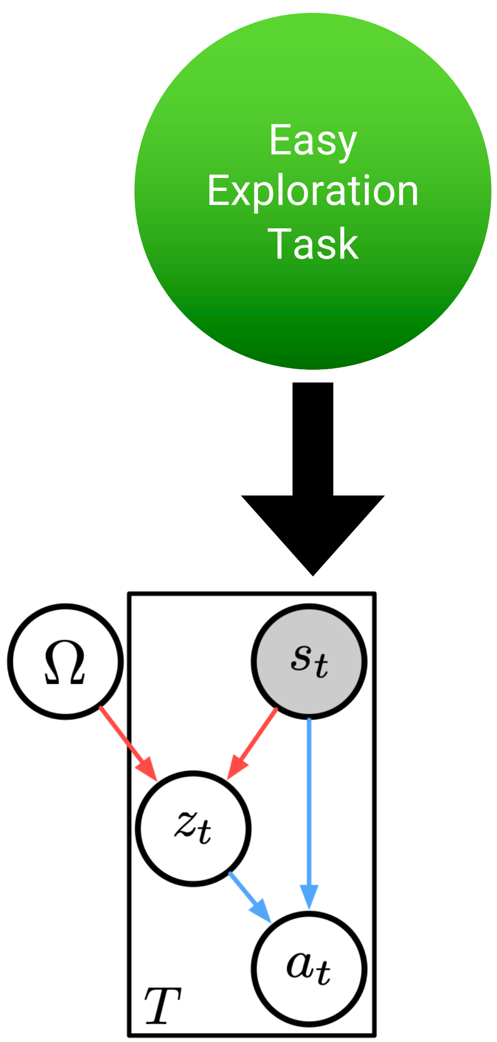
decoder
$p_{dec}(a_t|s_t, z_t)$



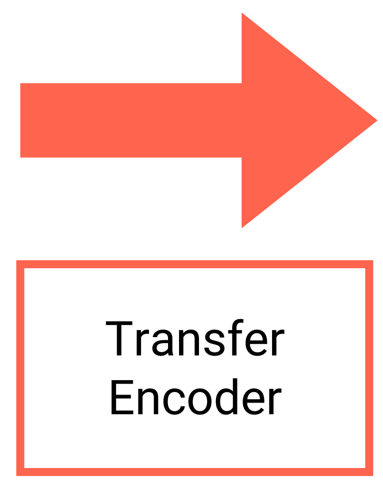
Transfer
Encoder

Sub-goal Transfer

Learn sub-goal detector
for simple task

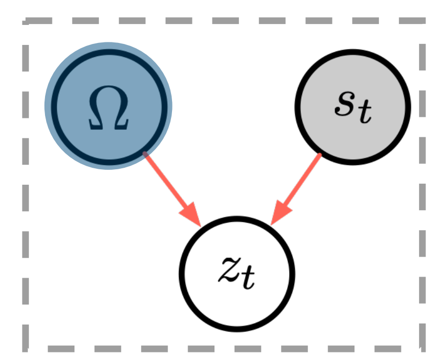


policy
$\pi(a_t|s_t,\Omega)$
encoder
$p_{enc}(z_t|s_t,\Omega)$
decoder
$p_{dec}\left(a_t|s_t,z_t\right)$



frozen sub-goal
detector

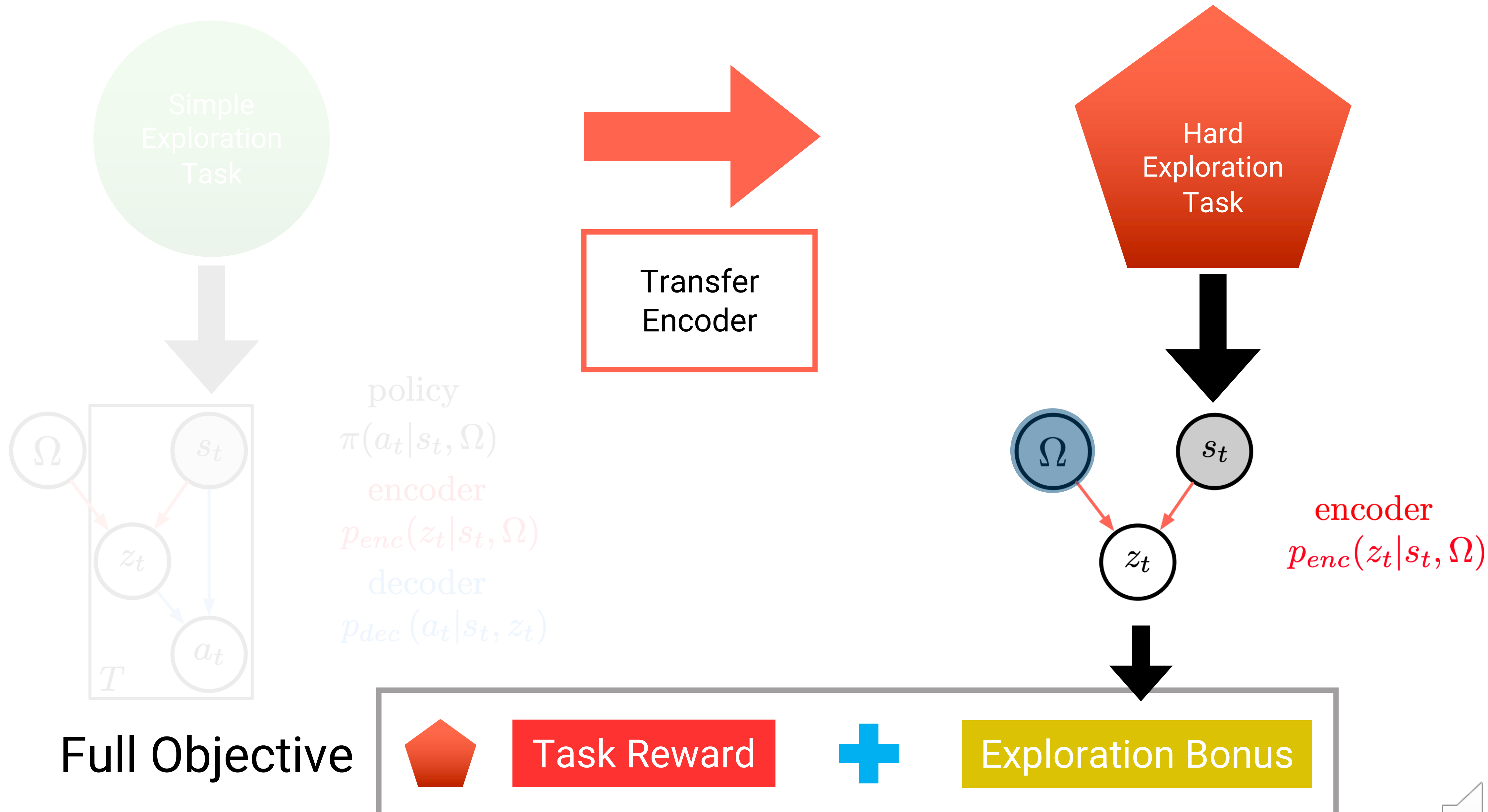
marginalize
omega



encoder
$p_{enc}(z_t|s_t,\Omega)$

Sub-goal Transfer

Learn sub-goal detector
for simple task



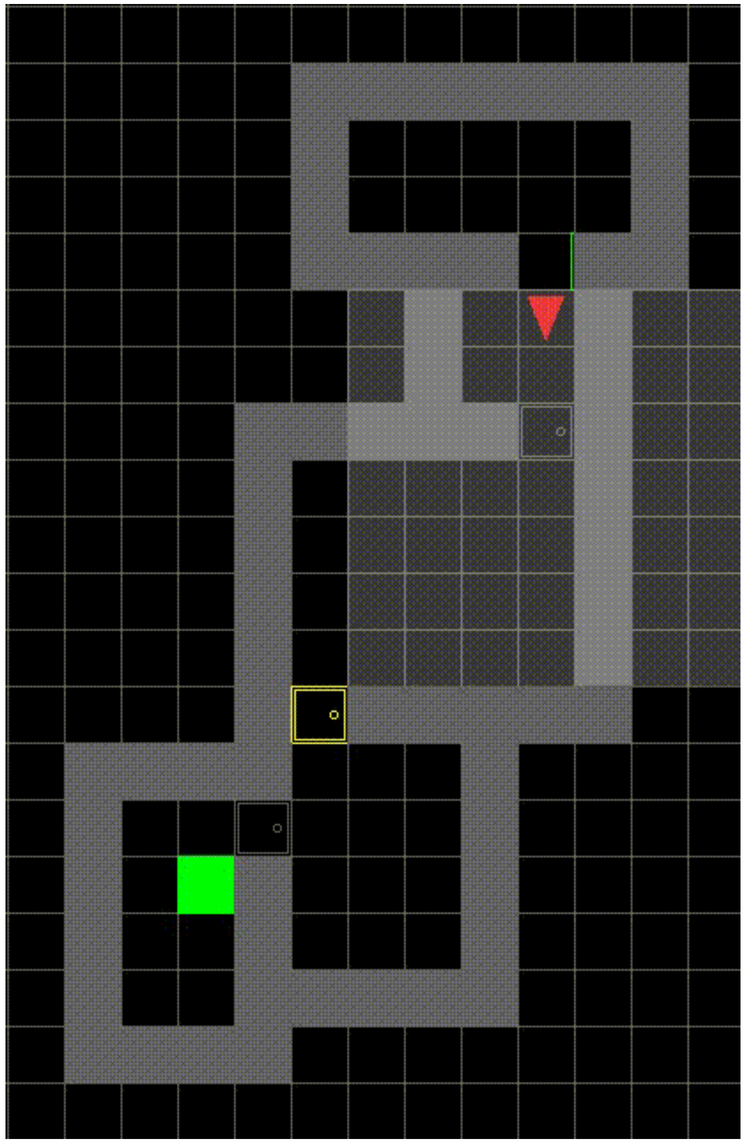
Experimental Evaluation

- Set of environments: **Gym-minigrid**

$$s_t \rightarrow \text{NxN Image}$$
$$G_t \rightarrow \text{Vector to Goal}$$
$$\mathcal{A} = \{\text{fwd, left, right, toggle}\}$$

- External Task: **Point-goal navigation (green square)**

- External Reward: **+1 (decaying) on goal reached**



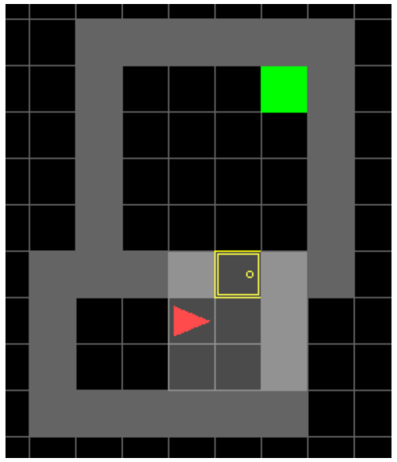
N: # of rooms
S: max room size

Gif Source: "Minimalistic Gridworld Environment for OpenAI Gym", 2018, github.com/maximecb/gym-minigrid

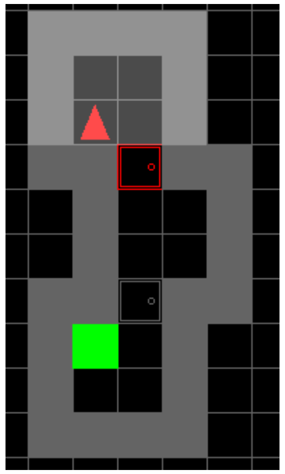
Easy Transfer

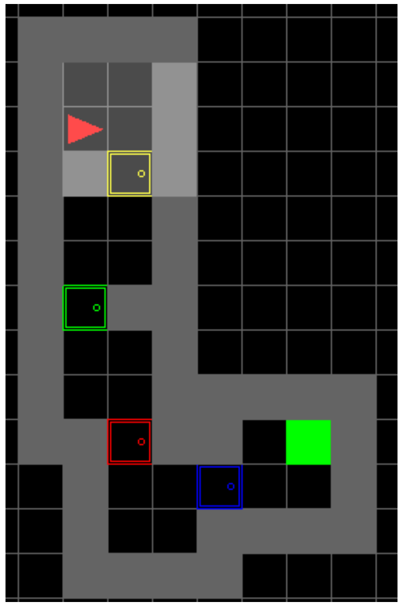
Transfer experiments from Goyal et al.[1]

N: # of rooms
S: max room size



Increasing
# of rooms





N2 S6

N3 S4

N3 S5

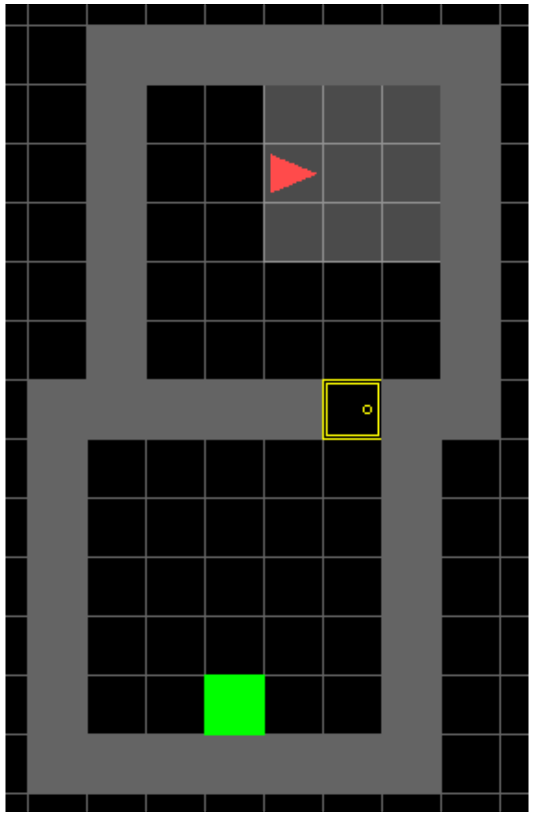
Source

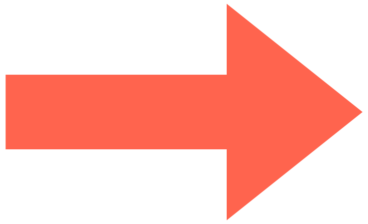
Target

Target

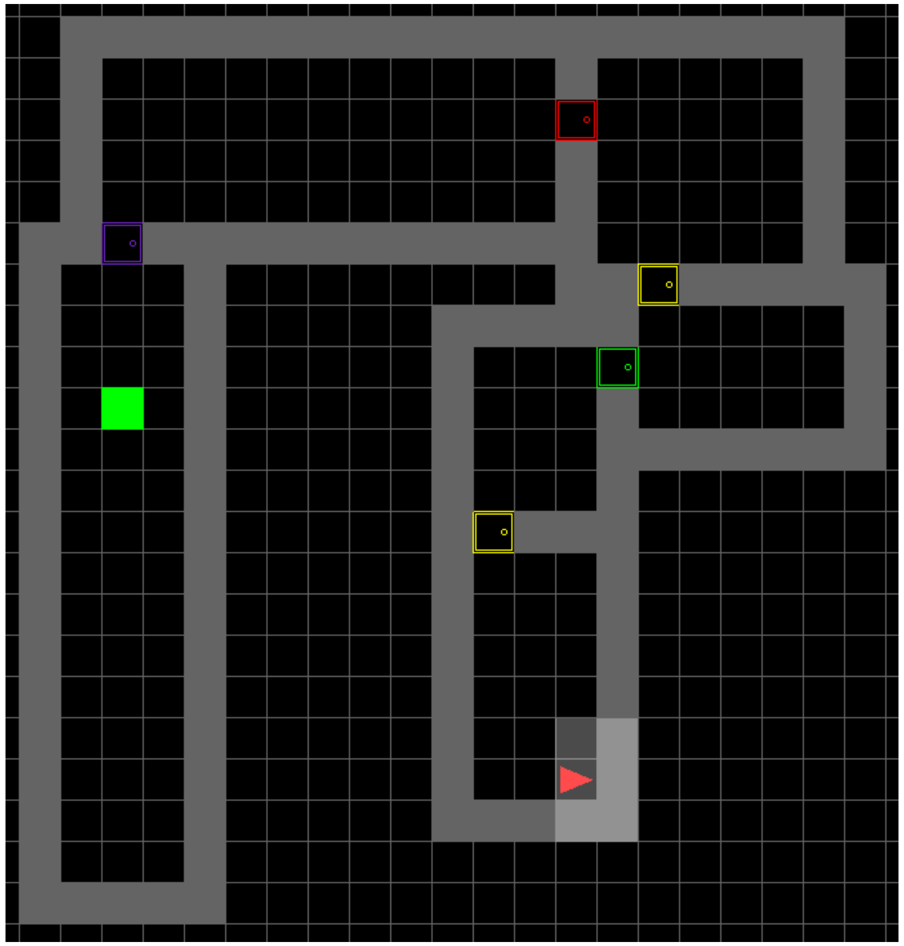
[1] Goyal, Anirudh, et al. "*Infobot: Transfer and exploration via the information bottleneck.*" arXiv preprint arXiv:1901.10902 (2019).

Challenging Transfer





Increasing
# of rooms
and max size



N2 S10

N6 S25

Success Evaluation

Success: % of times goal reached over 512 different environments
S.E.M.: Standard error of mean over 10 random seeds

Method	**MR-N3S4**	**MR-N5S4**	**MR-N6S25**
$p_\phi(Z_t \mid S_t, \Omega)$ pretrained on	MR-N2S6	MR-N2S6	MR-N2S10
InfoBot [Goyal *et al.*, 2019]	90%	85%	N/A
InfoBot (Our Implementation)	99.9%±0.1%	79.1%±11.6%	90.9%±1.2%
Count-based Baseline	99.7%±0.1%	99.7%±0.1%	86.8%±2.2%
DIAYN	99.7%±0.1%	95.4%±4.1%	0.1%±0.1%
Random Network	99.9%±0.1%	98.8%±0.7%	79.5%±5.2%
Heuristic Baseline	N/A	N/A	85.9%±3.0%
Ours ($\beta = 10^{-2}$)	99.3%±0.3%	99.4%±0.2%	92.9%±1.2%

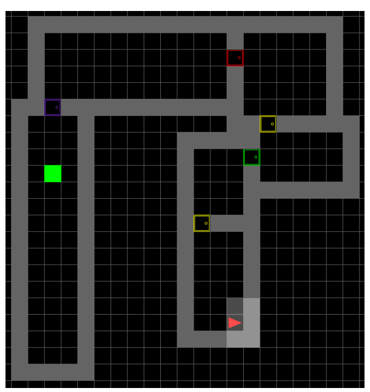
Success % ± s.e.m.

Easy Transfer

Challenging Transfer

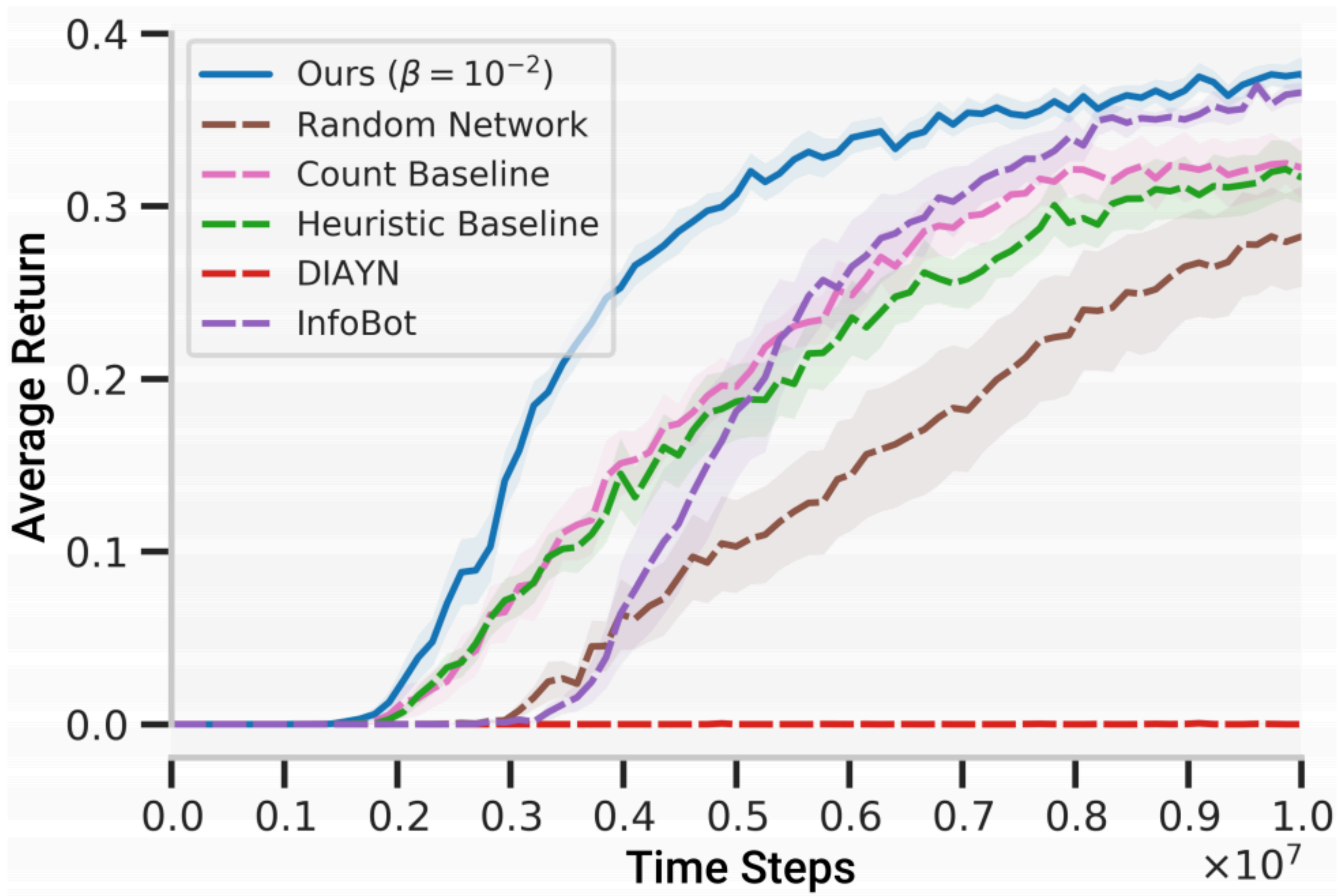
Average Return Evaluation

Sample Efficiency



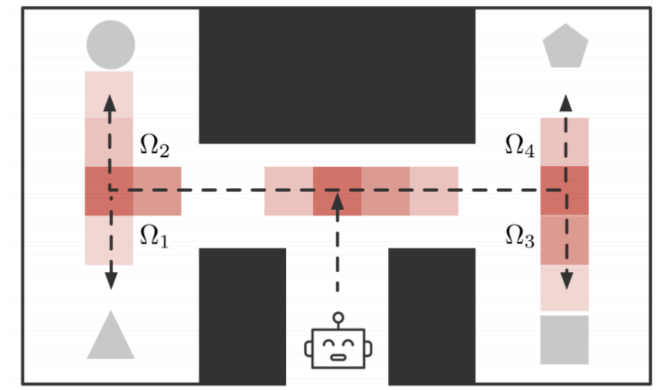
N = 6 rooms
S = 25 max room size

Challenging Transfer

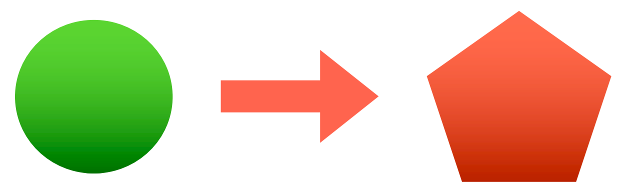


Summary

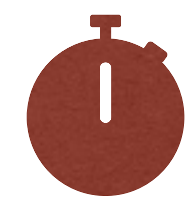
IR-VIC



- **Unsupervised objective** for sub-goal discovery
- **Transferable** sub-goals



- **Better exploration** and **sample efficiency** in hard exploration tasks



Visit our poster or watch our 20 minute video for more details!

Code (coming soon): github.com/nirbhayjm/irvic