A. Appendix

A.1. Proofs for Propositions and Theorems

Here we provide the proofs for the propositions and theorems used in our Option-GAIL.

A.1.1. PROOF FOR $\mathcal{O}_{\text{ONE-STEP}} = \mathcal{O}$

According to Assumption 1.1, an option can be activated at any state, thus the intra-option policy $\pi_o(a|s)$, break policy $\beta_{o'}(s)$ and inter-option policy $\pi_{\mathcal{O}}(o|s)$ are all well-defined on any $s \in \mathbb{S}, \forall o \in \mathcal{O}$. This suggests that $\pi_L(a|s,o) \equiv \pi_o(a|s)$ holds over all options on any state. For $\beta_{o'}(s)$, with Assumption 1.2, we have $\beta_{o'}(s) = 1 - \pi_H(o'|s,o')$ and for $\pi_{\mathcal{O}}(o|s)$ we have $\pi_{\mathcal{O}}(o|s) = \left.\frac{\pi_H(o|s,o')}{\sum_{o\neq o'}\pi_H(o|s,o')}\right|_{\forall o'\neq o} = \frac{\sum_{o'\neq o}\pi_H(o|s,o')}{\sum_{o'\neq o}\sum_{o\neq o'}\pi_H(o|s,o')}$. Also, with $o_{-1} \equiv \#$, it can be directly found that $\tilde{\mu}_0(s,o) = \tilde{\mu}_0(s, o=\#) \equiv \mu_0(s)$. Since $\mathbb{S}, \mathbb{A}, R_s^a, P_{s,s'}^a, \gamma$ are all defined the same between $\mathcal{O}_{\text{one-step}}$ and $\mathcal{O}$, we can get that$\mathcal{O}_{\text{one-step}} = \mathcal{O}$ holds under Assumption 1, and there exists an one-to-one mapping between $\big(\pi_H(o|s,o'), \pi_L(a|s,o)\big)$ and $\big(\pi_o(a|s), \beta_{o'}(s), \pi_{\mathcal{O}}(o|s)\big)$. $\square$

Combining with Theorem 1, this equivalency also suggests:

$$\rho_{\tilde{\pi}}(s,a,o,o') = \rho_{\tilde{\pi}^\star}(s,a,o,o') \Leftrightarrow \tilde{\pi} = \tilde{\pi}^\star \Leftrightarrow \big(\pi_o(a|s), \pi_{\mathcal{O}}(o|s), \beta_{o'}(s)\big) = \big(\pi_o^\star(a|s), \pi_{\mathcal{O}}^\star(o|s), \beta_{o'}^\star(s)\big). \tag{11}$$

A.1.2. PROOF FOR THEOREM 1

The proof of Theorem 1 can be derived similar as that from Syed et al. (2008) by defining an augmented MDP with options: $\tilde{s}_t \doteq (s_t, o_{t-1}) \in \mathbb{S} \times \mathbb{O}^+, \tilde{a}_t \doteq (a_t, o_t^A) \in \mathbb{A} \times \mathbb{O}, \tilde{\pi}(\tilde{a}_t|\tilde{s}_t) \doteq \pi_L(a_t|s_t, o_t^A)\pi_H(o_t^A|s_t, o_{t-1}), \tilde{P}_{\tilde{s}_t,\tilde{s}_{t+1}}^{\tilde{a}_t} \doteq P_{s_t,s_{t+1}}^{a_t}\mathbb{1}_{o_t=o_t^A}$, where we denote o_t used in $\tilde{a}_t$ as o_t^A for better distinguish from the option chosen in $\tilde{s}_{t+1}$, despite they should actually be the same.

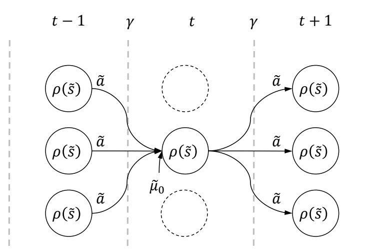

Figure 8. Illustration of the Bellman Flow on augmented MDP with options.

With the sugmented MDP, we can rewrite:

$$\begin{aligned}\rho(\tilde{s},\tilde{a}) &\doteq \rho(s,a,o,o') \\ &= \pi_L(a|s,o)\pi_L(o|s,o')\left(\tilde{\mu}_0(s,o') + \gamma\sum_{s',a',o''}\rho(s',a',o',o'')P_{s',s}^{a'}\right) \\ &= \tilde{\pi}(\tilde{a}|\tilde{s})\left(\tilde{\mu}_0 + \gamma\sum_{\tilde{s}',\tilde{a}'}\rho(\tilde{s}',\tilde{a}')\tilde{P}_{\tilde{s}',\tilde{s}}^{\tilde{a}'}\right)\end{aligned} \tag{12}$$

and construct a $\tilde{\pi}$-specific Bellman Flow constraint similar as that introduced by Syed et al. (2008):

$$\rho(\tilde{s},\tilde{a}) = \tilde{\pi}(\tilde{a}|\tilde{s})\left(\tilde{\mu}_0(\tilde{s}) + \gamma\sum_{\tilde{s}',\tilde{a}'}\rho(\tilde{s}',\tilde{a}')\tilde{P}_{\tilde{s}',\tilde{s}}^{\tilde{a}'}\right) \tag{13}$$

$$\rho(\tilde{s},\tilde{a}) \geq 0. \tag{14}$$

Now we build the relation between the option-occupancy measurement $\rho_{\tilde{\pi}}(\tilde{s}, \tilde{a})$ and the policy $\tilde{\pi}(\tilde{a}|\tilde{s})$.

Lemma 1 *The option-occupancy measurement of $\tilde{\pi}$ which is defined as $\rho_{\tilde{\pi}}(\tilde{s}, \tilde{a}) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathbb{1}_{(\tilde{s}_t=\tilde{s}, \tilde{a}_t=\tilde{a})}\right]$ satisfies the $\tilde{\pi}$-specific Bellman Flow constraint in Equation 13-14.*

proof: it can be directly find that Equation 14 is always satisfied as $\rho_{\tilde{\pi}}(\tilde{s}, \tilde{a}) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathbb{1}_{(\tilde{s}_t=\tilde{s}, \tilde{a}_t=\tilde{a})}\right] \geq 0$ always holds, we now verify the constraint in Equation 13:

$$
\begin{aligned}
\rho_{\tilde{\pi}}(\tilde{s}, \tilde{a}) &= \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathbb{1}_{(\tilde{s}_t=\tilde{s}, \tilde{a}_t=\tilde{a})}\right] = \sum_{t=0}^{\infty} \gamma^t P(\tilde{s}_t = \tilde{s}, \tilde{a}_t = \tilde{a}) && (15)\\
&= \tilde{\pi}(\tilde{a}|\tilde{s})\tilde{\mu}_0(\tilde{s}) + \sum_{t=1}^{\infty} \gamma^t P(\tilde{s}_t = \tilde{s}, \tilde{a}_t = \tilde{a}) && (16)\\
&= \tilde{\pi}(\tilde{a}|\tilde{s})\tilde{\mu}_0(\tilde{s}) + \sum_{t=1}^{\infty} \gamma^t \sum_{\tilde{s}', \tilde{a}'} P(\tilde{s}_t = \tilde{s}, \tilde{a}_t = \tilde{a}, \tilde{s}_{t-1} = \tilde{s}', \tilde{a}_{t-1} = \tilde{a}') && (17)\\
&= \tilde{\pi}(\tilde{a}|\tilde{s}) \left(\tilde{\mu}_0(\tilde{s}) + \sum_{t=1}^{\infty} \gamma^t \sum_{\tilde{s}', \tilde{a}'} \tilde{P}^{\tilde{a}'}_{\tilde{s}', \tilde{s}} P(\tilde{s}_{t-1} = \tilde{s}', \tilde{a}_{t-1} = \tilde{a}')\right) && (18)\\
&= \tilde{\pi}(\tilde{a}|\tilde{s}) \left(\tilde{\mu}_0(\tilde{s}) + \gamma \sum_{\tilde{s}', \tilde{a}'} \tilde{P}^{\tilde{a}'}_{\tilde{s}', \tilde{s}} \sum_{t=0}^{\infty} \gamma^t P(\tilde{s}_t = \tilde{s}', \tilde{a}_t = \tilde{a}')\right) && (19)\\
&= \tilde{\pi}(\tilde{a}|\tilde{s}) \left(\tilde{\mu}_0(\tilde{s}) + \gamma \sum_{\tilde{s}', \tilde{a}'} \tilde{P}^{\tilde{a}'}_{\tilde{s}', \tilde{s}} \mathbb{E}_{\tilde{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t \mathbb{1}_{(\tilde{s}_t=\tilde{s}', \tilde{a}_t=\tilde{a}')}\right]\right) && (20)\\
&= \tilde{\pi}(\tilde{a}|\tilde{s}) \left(\tilde{\mu}_0(\tilde{s}) + \gamma \sum_{\tilde{s}', \tilde{a}'} \rho_{\tilde{\pi}}(\tilde{s}', \tilde{a}') \tilde{P}^{\tilde{a}'}_{\tilde{s}', \tilde{s}}\right) \qquad \square && (21)
\end{aligned}
$$

Lemma 2 *The function that satisfies the $\tilde{\pi}$-specific Bellman Flow constraint in Equation 13-14 is unique.*

proof: we first define an operator for policy $\tilde{\pi}$: $\mathcal{T}^{\tilde{\pi}} : R^{|\mathbb{S}\times\mathbb{O}^+|\times|\mathbb{A}\times\mathbb{O}|} \mapsto R^{|\mathbb{S}\times\mathbb{O}^+|\times|\mathbb{A}\times\mathbb{O}|}$ for any function $f \in R^{|\mathbb{S}\times\mathbb{O}^+|\times|\mathbb{A}\times\mathbb{O}|}$: $\left(\mathcal{T}^{\tilde{\pi}} f\right)(\tilde{s}, \tilde{a}) \doteq \tilde{\pi}(\tilde{a}|\tilde{s}) \left(\tilde{\mu}_0(\tilde{s}) + \gamma \sum_{\tilde{s}', \tilde{a}'} f(\tilde{s}', \tilde{a}') \tilde{P}^{\tilde{a}'}_{\tilde{s}', \tilde{s}}\right)$, then for any two functions $\rho_1(\tilde{s}, \tilde{a}) \geq 0, \rho_2(\tilde{s}, \tilde{a}) \geq 0$ satisfy $\rho_1 = \mathcal{T}^{\tilde{\pi}} \rho_1, \rho_2 = \mathcal{T}^{\tilde{\pi}} \rho_2$, we have:

$$
\begin{aligned}
& \sum_{\tilde{s}, \tilde{a}} |\rho_1 - \rho_2| (\tilde{s}, \tilde{a}) = \sum_{\tilde{s}, \tilde{a}} \left|\mathcal{T}^{\tilde{\pi}} \rho_1 - \mathcal{T}^{\tilde{\pi}} \rho_2\right| (\tilde{s}', \tilde{a}') && (22)\\
= \; & \sum_{\tilde{s}, \tilde{a}} \left|\tilde{\pi}(\tilde{a}|\tilde{s}) \gamma \sum_{\tilde{s}', \tilde{a}'} \tilde{P}^{\tilde{a}'}_{\tilde{s}', \tilde{s}} (\rho_1 - \rho_2)(\tilde{s}', \tilde{a}')\right| = \gamma \sum_{\tilde{s}, \tilde{a}} \left|\sum_{\tilde{s}', \tilde{a}'} p(\tilde{s}, \tilde{a}|\tilde{s}', \tilde{a}') (\rho_1 - \rho_2)(\tilde{s}', \tilde{a}')\right| && (23)\\
\leq \; & \gamma \sum_{\tilde{s}, \tilde{a}} \sum_{\tilde{s}', \tilde{a}'} p(\tilde{s}, \tilde{a}|\tilde{s}', \tilde{a}') |\rho_1 - \rho_2| (\tilde{s}', \tilde{a}') = \gamma \sum_{\tilde{s}', \tilde{a}'} |\rho_1 - \rho_2| (\tilde{s}', \tilde{a}') && (24)\\
= \; & \gamma \sum_{\tilde{s}, \tilde{a}} |\rho_1 - \rho_2| (\tilde{s}, \tilde{a}) && (25)\\
\because \; & \sum_{\tilde{s}, \tilde{a}} |\rho_1 - \rho_2| (\tilde{s}, \tilde{a}) \geq 0, \gamma < 1 && (26)\\
\therefore \; & \sum_{\tilde{s}, \tilde{a}} |\rho_1 - \rho_2| (\tilde{s}, \tilde{a}) = 0 \Rightarrow \rho_1 = \rho_2 \qquad \square && (27)
\end{aligned}
$$

Lemma 3 *There is a bijection between $\tilde{\pi}(\tilde{a}|\tilde{s})$ and $\left(\pi_H(o|s,o'), \pi_L(a|s,o)\right)$, where $\tilde{\pi}(\tilde{a}|\tilde{s}) = \tilde{\pi}(a,o|s,o') = \pi_L(a|s,o)\pi_H(o|s,o')$ and $\pi_H(o|s,o') = \sum_a \tilde{\pi}(a,o|s,o'), \pi_L(a|s,o) = \frac{\tilde{\pi}(a,o|s,o')}{\sum_a \tilde{\pi}(a,o|s,o')}\Big|_{\forall o'} = \frac{\sum_{o'} \tilde{\pi}(a,o|s,o')}{\sum_{a,o'} \tilde{\pi}(a,o|s,o')}$*

With Lemma 1 and Lemma 2, the proof of Theorem 1 is provided:

proof: For any $\rho(s,a,o,o') = \rho(\tilde{s},\tilde{a}) \in \mathbb{D} = \left\{\rho(\tilde{s},\tilde{a}) \geq 0; \sum_{\tilde{a}} \rho(\tilde{s},\tilde{a}) = \tilde{\mu}_0(\tilde{s}) + \gamma \sum_{\tilde{s}',\tilde{a}'} \rho(\tilde{s}',\tilde{a}')\tilde{P}^{\tilde{a}'}_{\tilde{s}',\tilde{s}}\right\}$, and a policy $\tilde{\pi}(\tilde{a}|\tilde{s})$ satisfies:

$$\tilde{\pi}(\tilde{a}|\tilde{s}) = \frac{\rho(\tilde{s},\tilde{a})}{\sum_{\tilde{a}} \rho(\tilde{s},\tilde{a})} = \frac{\rho(\tilde{s},\tilde{a})}{\tilde{\mu}_0(\tilde{s}) + \gamma \sum_{\tilde{s}',\tilde{a}'} \rho(\tilde{s}',\tilde{a}')\tilde{P}^{\tilde{a}'}_{\tilde{s}',\tilde{s}}},. \tag{28}$$

With Equation 28 ρ should be a solution of Equation 13-14, and with Lemma 1-2, the solution is unique and equals to the occupancy measurement of $\tilde{\pi}$. With Lemma 3, ρ is also the unique occupancy measurement of (π_H, π_L).

On the other hand, If $\rho_{\tilde{\pi}}$ is the occupancy measurement of $\tilde{\pi}$, we have:

$$\sum_{\tilde{a}} \tilde{\pi}(\tilde{a}|\tilde{s}) = 1 = \frac{\sum_{\tilde{a}} \rho_{\tilde{\pi}}(\tilde{s},\tilde{a})}{\tilde{\mu}_0(\tilde{s}) + \gamma \sum_{\tilde{s}',\tilde{a}'} \rho_{\tilde{\pi}}(\tilde{s}',\tilde{a}')\tilde{P}^{\tilde{a}'}_{\tilde{s}',\tilde{s}}}, \tag{29}$$

which indicates that $\rho_{\tilde{\pi}} \in \mathbb{D}$ and $\tilde{\pi}(a,o|s,o') = \frac{\rho_{\tilde{\pi}}(s,a,o,o')}{\sum_{a,o} \rho_{\tilde{\pi}}(s,a,o,o')}$, also:

$$\pi_H(o|s,o') = \sum_a \tilde{\pi}(a,o|s,o') = \frac{\sum_a \rho_{\tilde{\pi}}(s,a,o,o')}{\sum_{a,o} \rho_{\tilde{\pi}}(s,a,o,o')} \tag{30}$$

$$\pi_L(a|s,o) = \frac{\sum_{o'} \tilde{\pi}(a,o|s,o')}{\sum_{a,o'} \tilde{\pi}(a,o|s,o')} = \frac{\sum o' \rho_{\tilde{\pi}}(s,a,o,o')}{\sum_{a,o'} \rho_{\tilde{\pi}}(s,a,o,o')} \quad \square \tag{31}$$

A.1.3. PROOF FOR THEOREM 2

We first adapt the corollary on Ghasemipour et al. (2020) into its option-version.

Lemma 4 *Optimizing the f-divergence between $\rho_{\tilde{\pi}}$ and $\rho_{\tilde{\pi}_E}$ equals to perform $\tilde{\pi}^\star = HRL(c^\star)$ with $c^\star = HIRL_\psi(\tilde{\pi}_E)$: $\tilde{\pi}^\star = HRL \circ HIRL_\psi(\tilde{\pi}_E) = \arg\min_{\tilde{\pi}} -\mathbb{H}(\tilde{\pi}) + D_f\left(\rho_{\tilde{\pi}}(s,a,o,o') \| \rho_{\tilde{\pi}_E}(s,a,o,o')\right)$*

proof: we take similar deviations from that provided by Ghasemipour et al. (2020). Let f be a function defining a f-divergence and let $f^\star$ be the convex conjugate of f. Given $\rho_{\tilde{\pi}_E}$ and cost functions $c(s,a,o,o')$ defined on $\mathbb{S} \times \mathbb{A} \times \mathbb{O} \times \mathbb{O}^+$, we can define the cost function regularizer used by our option-based HIRL as $\psi_f(c) \doteq$

$\mathbb{E}_{\rho_{\tilde{\pi}_E}(s,a,o,o')}\left[f^\star\left(c(s,a,o,o')\right) - c(s,a,o,o')\right]$ and a similar relation holds:

$$
\begin{aligned}
& \psi_f^\star\left(\rho_{\tilde{\pi}}(s,a,o,o') - \rho_{\tilde{\pi}_E}(s,a,o,o')\right) && (32)\\
= & \sup_{c()}\left[\sum_{s,a,o,o'}(\rho_{\tilde{\pi}} - \rho_{\tilde{\pi}_E})(s,a,o,o')c(s,a,o,o') - \psi_f(c)\right] && (33)\\
= & \sup_{c()}\left[\sum_{s,a,o,o'}(\rho_{\tilde{\pi}} - \rho_{\tilde{\pi}_E})(s,a,o,o')c(s,a,o,o')\right.\\
& \left. - \sum_{s,a,o,o'}\rho_{\tilde{\pi}_E}(s,a,o,o')\left(f^\star\left(c(s,a,o,o')\right) - c(s,a,o,o')\right)\right] && (34)\\
= & \sup_{c()}\left[\sum_{s,a,o,o'}\left[\rho_{\tilde{\pi}}(s,a,o,o')c(s,a,o,o') - \rho_{\tilde{\pi}_E}(s,a,o,o')f^\star\left(c(s,a,o,o')\right)\right]\right] && (35)\\
= & \sup_{c()}\left[\mathbb{E}_{\rho_{\tilde{\pi}}}\left[c(s,a,o,o')\right] - \mathbb{E}_{\rho_{\tilde{\pi}_E}}\left[f^\star\left(c(s,a,o,o')\right)\right]\right], \text{ let } T_\omega = c && (36)\\
= & \sup_{T_\omega}\left[\mathbb{E}_{\rho_{\tilde{\pi}}}\left[T_\omega(s,a,o,o')\right] - \mathbb{E}_{\rho_{\tilde{\pi}_E}}\left[f^\star\left(T_\omega(s,a,o,o')\right)\right]\right] && (37)\\
= & D_f\left(\rho_{\tilde{\pi}}(s,a,o,o') \| \rho_{\tilde{\pi}_E}(s,a,o,o')\right), && (38)
\end{aligned}
$$

where $\tilde{\pi}^\star = \text{HRL} \circ \text{HIRL}_\psi(\tilde{\pi}_E) = \arg\min_{\tilde{\pi}} -\mathbb{H}(\tilde{\pi}) + \psi_f^\star\left(\rho_{\tilde{\pi}}(s,a,o,o') - \rho_{\tilde{\pi}_E}(s,a,o,o')\right) = \arg\min_{\tilde{\pi}} -\mathbb{H}(\tilde{\pi}) + D_f\left(\rho_{\tilde{\pi}}(s,a,o,o') \| \rho_{\tilde{\pi}_E}(s,a,o,o')\right)$. □

Similar as Ghasemipour et al. (2020), we omit the entropy regularizer term in Lemma 4, thus after the optimization in M-step we have $D_f\left(\rho_{\tilde{\pi}^{n-1}}(s,a,o,o') \| \rho_E(s,a)[p_{\tilde{\pi}^n}(o,o'|s,a)\right) \geq D_f\left(\rho_{\tilde{\pi}^n}(s,a,o,o') \| \rho_E(s,a)p_{\tilde{\pi}^n}(o,o'|s,a)\right)$. Now we are ready for proving Theorem 2:

proof: Since the option of expert is inferred based on the policy $\tilde{\pi}^n$ on each optimization step, we separate the expert option-occupancy measurement estimated with $\tilde{\pi}^n$ as: $\rho_{\tilde{\pi}_E}(s,a,o,o') = \rho_E(s,a)p_{\tilde{\pi}^n}(o,o'|s,a)$. By repeating the definition of Q_n in our main paper, we have

$$
\begin{aligned}
Q_n = & \mathbb{E}_{p_{\tilde{\pi}^{n-1}}(o,o'|s,a)}\left[D_f\left(\rho_{\tilde{\pi}^n}(s,a,o,o') \| \rho_{\tilde{\pi}_E}(s,a,o,o')\right)\right] && (39)\\
= & \sum_{s,a,o,o'}\rho_E(s,a)p_{\tilde{\pi}^{n-1}}(o,o'|s,a)f\left(\frac{\rho_{\tilde{\pi}^n}(s,a,o,o')}{\rho_E(s,a)p_{\tilde{\pi}^{n-1}}(o,o'|s,a)}\right)\\
\geq & \sum_{s,a}\rho_E(s,a)f\left(\frac{\rho_{\tilde{\pi}^n}(s,a)}{\rho_E(s,a)}\right) \quad (f \text{ is convex}) && (40)\\
= & \sum_{s,a,o,o'}\rho_E(s,a)p_{\tilde{\pi}^n}(o,o'|s,a)f\left(\frac{\rho_{\tilde{\pi}^n}(s,a,o,o')}{\rho_E(s,a)p_{\tilde{\pi}^n}(o,o'|s,a)}\right) \quad \text{(E-Step)}\\
\geq & \sum_{s,a,o,o'}\rho_E(s,a)p_{\tilde{\pi}^n}(o,o'|s,a)f\left(\frac{\rho_{\tilde{\pi}^{n+1}}(s,a,o,o')}{\rho_E(s,a)p_{\tilde{\pi}^n}(o,o'|s,a)}\right) \quad \text{(M-Step)}\\
= & Q_{n+1}. \quad \square && (41)
\end{aligned}
$$

With Equation 39, Equation 40 and Equation 41 we can also obtain:

$$
\begin{aligned}
& D_f\left(\rho_{\tilde{\pi}^n}(s,a,o,o') \| \rho_E(s,a)p_{\tilde{\pi}^n}(o,o'|s,a)\right) \geq D_f\left(\rho_{\tilde{\pi}^{n+1}}(s,a,o,o') \| \rho_E(s,a)p_{\tilde{\pi}^{n+1}}(o,o'|s,a)\right) && (42)\\
\Rightarrow \quad & D_f\left(\rho_{\tilde{\pi}^n}(s,a) \| \rho_E(s,a)\right) \geq D_f\left(\rho_{\tilde{\pi}^{n+1}}(s,a) \| \rho_E(s,a)\right) && (43)
\end{aligned}
$$

A.2. Experimental Details and Extra Results

Here we provide more comparative results on several counterparts, as well as the experimental details.[2].

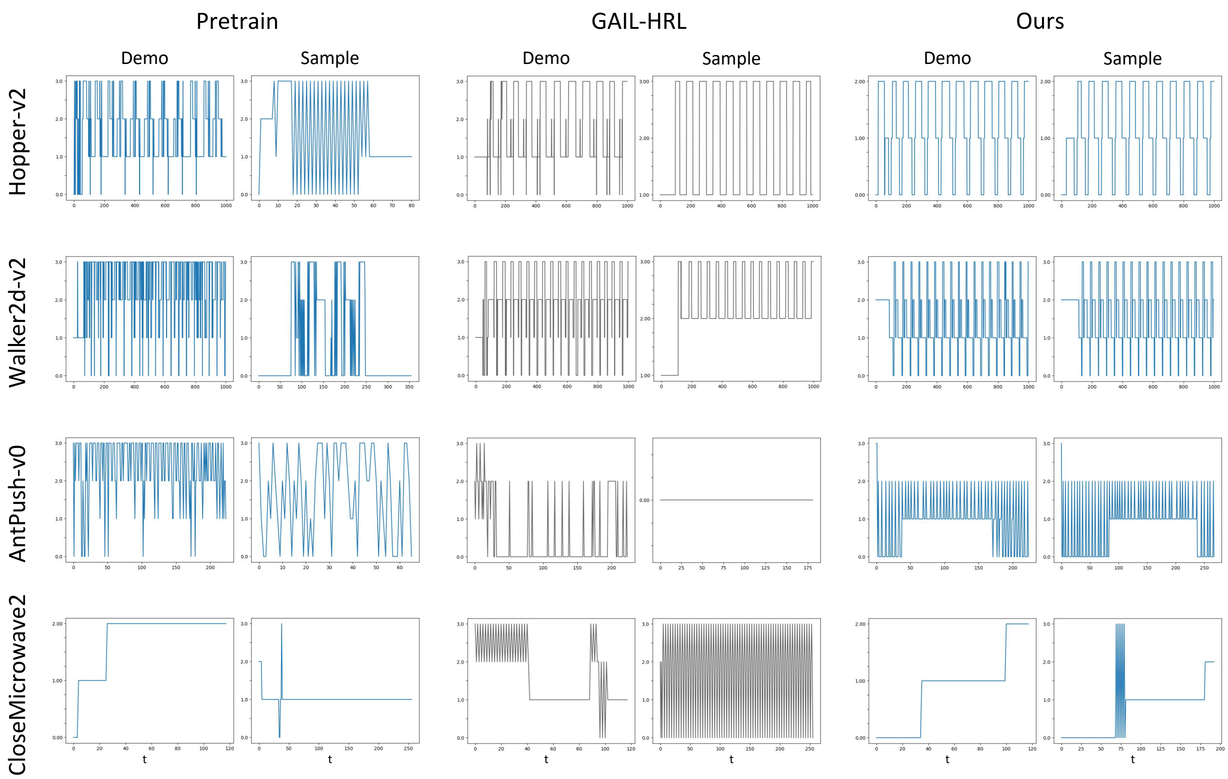

Figure 9. Visualization of the options activated at each step, learned respectively by pretraining and fixing high-level policy(Pretrain, refers to Directed-info GAIL (Sharma et al., 2018)), Mixer of Expert(MoE, refers to OptionGAN (Henderson et al., 2018)), GAIL-HRL and our proposed method. 'Demo' denotes the options inferred from the expert, and 'Sample' denotes the options used by agent when doing self-explorations. The effectiveness of our proposed method on regularizing the option switching is obvious by comparing the consistent switching tendencies between Demo and Sample.

A.2.1. Extra Results

Table 2. Comparative results. All results are measured by the average **maximum average reward-sum** among different trails.

	Hopper-v2	Walker2d-v2	AntPush-v0	CloseMicrowave2
Demos $(s, a) \times T$	$(\mathbb{R}^{11}, \mathbb{R}^{3}) \times 1$k	$(\mathbb{R}^{17}, \mathbb{R}^{6}) \times 5$k	$(\mathbb{R}^{107}, \mathbb{R}^{8}) \times 50$k	$(\mathbb{R}^{101}, \mathbb{R}^{8}) \times 1$k
Demo Reward	3656.17±0.0	5005.80±36.18	116.60±14.07	—
GAIL	535.29±7.19	2787.87±2234.46	56.45±3.17	39.14±12.87
Pretrain	436.55±27.74	891.70±100.58	-0.07±1.50	74.34±20.16
MoE	3254.12±446.78	2722.11±2217.80	39.73±37.00	33.33±25.07
GAIL-HRL	3697.40±1.14	3687.63±982.99	20.53±6.90	56.95±25.74
Ours	**3700.42±1.70**	**4836.85±100.09**	**95.00±2.70**	**100.74±21.33**

A.2.2. Experimental Details

[2]The source code is provided at Option-GAIL.git. For setting up the environments correctly, please also refer to OpenAI-Gym (Brockman et al., 2016) and RLBench (James et al., 2020)

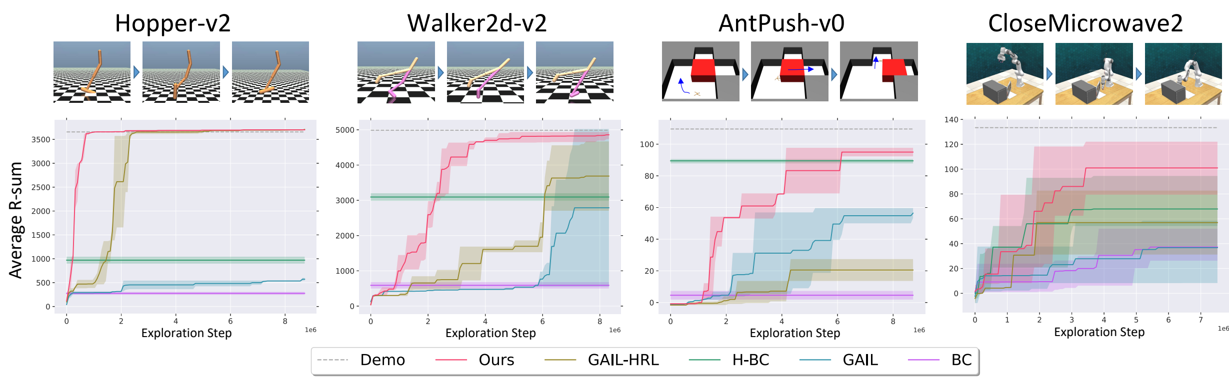

Figure 10. comparison of learning performance on four environments. We compare the **maximum average reward-sums** vs. exploration steps on different environments. The solid line indicates the average performance among several trials under different random seeds, while the shade indicates the range of the maximum average reward-sums over different trials.

$\|O\|$	Option-Viterbi / total (s)	%
2	$0.0938/57.785$	0.16%
3	$0.0884/90.199$	0.10%
4	$0.0840/102.00$	0.08%
5	$0.0938/126.05$	0.07%
6	$0.1014/142.64$	0.07%

Table 3. The computation time of Option-Viterbi comparing with the overall learning time costs

Table 4. Configurations and hyper-parameters

Name	Value	Name	Value
γ	0.99	learning rate	0.0003
$\lambda_{\mathbb{M}_L}$	0	$\lambda_{\mathbb{M}^H}$	0.01
batch size(T)	4096	mini batch size	64