

811 A Details of the Derivation

812 A.1 Derivations of Eq. (9) and Eq. (10)

813 In this section, we provide detailed derivations of Eq. (9) and Eq. (10). Given the expert sample
814 $\tau^0 \sim p_E$, the reverse process $p_\theta(\tau^{i-1}|\tau^i)$, and the corresponding forward process $q(\tau^{1:N}|\tau^0)$,

$$\begin{aligned}
& \mathbb{E}_{\tau^0 \sim p_E} \left[\log \sigma \left(\mathbb{E}_{\tau^{1:N} \sim q(\tau^{1:N}|\tau^0)} \log \frac{\prod_{i=1}^N p_\theta(\tau^{i-1}|\tau^i)}{\prod_{i=1}^N p_{\theta^{\text{old}}}(\tau^{i-1}|\tau^i)} \right) \right] \\
&= \mathbb{E}_{\tau^0 \sim p_E} \left[\log \sigma \left(\mathbb{E}_{\tau^{1:N} \sim q(\tau^{1:N}|\tau^0)} \sum_{i=1}^N \log \frac{p_\theta(\tau^{i-1}|\tau^i)}{p_{\theta^{\text{old}}}(\tau^{i-1}|\tau^i)} \right) \right] \\
&= \mathbb{E}_{\tau^0 \sim p_E} \left[\log \sigma \left(\sum_{i=1}^N \mathbb{E}_{\tau^{1:N} \sim q(\tau^{1:N}|\tau^0)} \log \frac{p_\theta(\tau^{i-1}|\tau^i)}{p_{\theta^{\text{old}}}(\tau^{i-1}|\tau^i)} \right) \right] \\
&= \mathbb{E}_{\tau^0 \sim p_E} \left[\log \sigma \left(\sum_{i=1}^N \mathbb{E}_{\tau^{i-1}, \tau^i \sim q(\tau^i|\tau^0)q(\tau^{i-1}|\tau^i, \tau^0)} \log \frac{p_\theta(\tau^{i-1}|\tau^i)}{p_{\theta^{\text{old}}}(\tau^{i-1}|\tau^i)} \right) \right] \\
&\geq \mathbb{E}_{\tau^0 \sim p_E} \left[\mathbb{E}_{i \sim \mathcal{U}(1, N), \tau^i \sim q(\tau_i|\tau_0)} \log \sigma \left(N \mathbb{E}_{\tau^{i-1} \sim q(\tau^{i-1}|\tau^i, \tau^0)} \log \frac{p_\theta(\tau^{i-1}|\tau^i)}{p_{\theta^{\text{old}}}(\tau^{i-1}|\tau^i)} \right) \right] \quad (\text{concavity}) \\
&= \mathbb{E}_{\tau^0 \sim p_E} \left[\mathbb{E}_{i \sim \mathcal{U}(1, N), \tau^i \sim q(\tau_i|\tau_0)} \log \sigma \left(N \mathbb{E}_{\tau^{i-1} \sim q(\tau^{i-1}|\tau^i, \tau^0)} \log \frac{q(\tau^{i-1}|\tau^i, \tau^0)}{p_{\theta^{\text{old}}}(\tau^{i-1}|\tau^i)} - \log \frac{q(\tau^{i-1}|\tau^i, \tau^0)}{p_\theta(\tau^{i-1}|\tau^i)} \right) \right] \quad (13)
\end{aligned}$$

815 The posterior $q(\tau^{i-1}|\tau^i, \tau^0)$ is tractable via Bayes' rule, since both $q(\tau^{i-1}|\tau^0)$ and $q(\tau^i|\tau^{i-1}, \tau^0) =$
816 $q(\tau^i|\tau^{i-1})$ are Gaussian distributions. The posterior $q(\tau^{i-1}|\tau^i, \tau^0)$ can be written as the following:

$$q(\tau^{i-1}|\tau^i, \tau_0) = \mathcal{N}(\tilde{\mu}_i(\tau^i, \tau_0), \tilde{\beta}_i I)$$

817 where $\tilde{\mu}_i(\tau^i, \tau_0) := \frac{1}{\sqrt{1-\beta_i}} \left(\tau^i - \frac{\beta_i}{\sqrt{1-\alpha_i}} z(\tau^i, \tau_0) \right)$, $z(\tau^i, \tau_0) := \frac{\tau^i - \sqrt{\alpha_i} \tau^0}{\sqrt{1-\alpha_i}} = \epsilon$. If we parameter-
818 ize the model p_θ as $\mathcal{N}(\mu_\theta(\tau^i, i), \sigma_i^2 I)$, then the KL divergence between these isotropic Gaussians
819 reduces to the squared error between the means: $\|\tilde{\mu}_i(\tau^i, \tau_0) - \mu_\theta(\tau_i, i)\|^2$. Ho et al. [15] proposes
820 to parameterize μ_θ using a predicted noise ϵ_θ , such as:

$$\mu_\theta(\tau^i, i) = \frac{1}{\sqrt{1-\beta_i}} \left(\tau^i - \frac{\beta_i}{\sqrt{1-\alpha_i}} \epsilon_\theta(\tau^i, i) \right)$$

821 Substituting this into the squared error gives:

$$\begin{aligned}
\left\| \tilde{\mu}_i(\tau^i, \tau_0) - \mu_\theta(\tau^i, i) \right\|^2 &= \left\| \frac{1}{\sqrt{1-\beta_i}} \left(\tau^i - \frac{\beta_i}{\sqrt{1-\alpha_i}} z(\tau^i, \tau_0) \right) - \mu_\theta(\tau^i, i) \right\|^2 \\
&= \left\| \frac{1}{\sqrt{1-\beta_i}} \left(\tau^i - \frac{\beta_i}{\sqrt{1-\alpha_i}} \epsilon \right) - \frac{1}{\sqrt{1-\beta_i}} \left(\tau^i - \frac{\beta_i}{\sqrt{1-\alpha_i}} \epsilon_\theta(\tau^i, i) \right) \right\|^2 \\
&= \frac{\beta_i}{(1-\beta_i)(1-\alpha_i)} \left\| \epsilon - \epsilon_\theta(\tau^i, i) \right\|^2
\end{aligned}$$

822 Therefore, minimizing the KL divergence is equivalent to minimizing $\|\epsilon - \epsilon_\theta(\tau^i, i)\|^2$. Substituting
823 this into Eq. 13 leads to the following equation:

$$\therefore \mathbb{E}_{\tau^0, i, \epsilon} \left[\log \sigma \left(N \cdot C_i \left(\|\epsilon - \epsilon_{\theta^{\text{old}}}(\tau^i, i)\|^2 - \|\epsilon - \epsilon_\theta(\tau^i, i)\|^2 \right) \right) \right].$$

824 Here, $C_i = \frac{\beta_i}{(1-\beta_i)(1-\alpha_i)}$, $\epsilon \sim \mathcal{N}(0, I)$, and $\tau^i \sim q(\tau^i|\tau^0)$, thus $\tau^i = \sqrt{\alpha_i} \tau^0 + (1-\alpha_i)\epsilon$. Follow-
825 ing Wallace et al. [32], we consider the weight $N \cdot C_i$ as a fixed constant value over i in practical
826 implementation. Similarly, given the generative sample $\bar{\tau}^0 \sim p_{\theta^{\text{old}}}$, the reverse process $p_{\theta^{\text{old}}}(\bar{\tau}^{i-1}|\bar{\tau}^i)$,
827 and the corresponding forward process $q_{\theta^{\text{old}}}(\bar{\tau}^{1:N}|\bar{\tau}^0) = q(\bar{\tau}^{1:N}|\bar{\tau}^0)$ (since they have the same

828 variance schedule),

$$\begin{aligned}
& \mathbb{E}_{\bar{\tau}^0 \sim p_{\theta^{\text{old}}}} \left[\log \sigma \left(\mathbb{E}_{\bar{\tau}^{1:N} \sim q(\tau^{1:N} | \bar{\tau}^0)} \log \frac{\prod_{i=1}^N p_{\theta^{\text{old}}}(\bar{\tau}^{i-1} | \bar{\tau}^i)}{\prod_{i=1}^N p_{\theta}(\bar{\tau}^{i-1} | \bar{\tau}^i)} \right) \right] \\
&= \mathbb{E}_{\bar{\tau}^0 \sim p_{\theta^{\text{old}}}} \left[\log \sigma \left(\mathbb{E}_{\bar{\tau}^{1:N} \sim q(\bar{\tau}^{1:N} | \bar{\tau}^0)} \sum_{i=1}^N \log \frac{p_{\theta^{\text{old}}}(\bar{\tau}^{i-1} | \bar{\tau}^i)}{p_{\theta}(\bar{\tau}^{i-1} | \bar{\tau}^i)} \right) \right] \\
&= \mathbb{E}_{\bar{\tau}^0 \sim p_{\theta^{\text{old}}}} \left[\log \sigma \left(\sum_{i=1}^N \mathbb{E}_{\bar{\tau}^{1:N} \sim q(\bar{\tau}^{1:N} | \bar{\tau}^0)} \log \frac{p_{\theta^{\text{old}}}(\bar{\tau}^{i-1} | \bar{\tau}^i)}{p_{\theta}(\bar{\tau}^{i-1} | \bar{\tau}^i)} \right) \right] \\
&= \mathbb{E}_{\bar{\tau}^0 \sim p_{\theta^{\text{old}}}} \left[\log \sigma \left(\sum_{i=1}^N \mathbb{E}_{\bar{\tau}^{i-1}, \bar{\tau}^i \sim q(\bar{\tau}^i | \bar{\tau}^0) q(\bar{\tau}^{i-1} | \bar{\tau}^i, \bar{\tau}^0)} \log \frac{p_{\theta^{\text{old}}}(\bar{\tau}^{i-1} | \bar{\tau}^i)}{p_{\theta}(\bar{\tau}^{i-1} | \bar{\tau}^i)} \right) \right] \\
&\geq \mathbb{E}_{\bar{\tau}^0 \sim p_{\theta^{\text{old}}}} \left[\mathbb{E}_{i \sim \mathcal{U}(1, N), \bar{\tau}^i \sim q(\bar{\tau}^i | \bar{\tau}^0)} \log \sigma \left(N \mathbb{E}_{\bar{\tau}^{i-1} \sim q(\bar{\tau}^{i-1} | \bar{\tau}^i, \bar{\tau}^0)} \log \frac{p_{\theta^{\text{old}}}(\bar{\tau}^{i-1} | \bar{\tau}^i)}{p_{\theta}(\bar{\tau}^{i-1} | \bar{\tau}^i)} \right) \right] \quad (\text{concavity}) \\
&= \mathbb{E}_{\bar{\tau}^0 \sim p_{\theta^{\text{old}}}} \left[\mathbb{E}_{i \sim \mathcal{U}(1, N), \bar{\tau}^i \sim q(\bar{\tau}^i | \bar{\tau}^0)} \log \sigma \left(N \mathbb{E}_{\bar{\tau}^{i-1} \sim q(\bar{\tau}^{i-1} | \bar{\tau}^i, \bar{\tau}^0)} \log \frac{q(\bar{\tau}^{i-1} | \bar{\tau}^i, \bar{\tau}^0)}{p_{\theta}(\bar{\tau}^{i-1} | \bar{\tau}^i)} - \log \frac{q(\bar{\tau}^{i-1} | \bar{\tau}^i, \bar{\tau}^0)}{p_{\theta^{\text{old}}}(\bar{\tau}^{i-1} | \bar{\tau}^i)} \right) \right] \\
&\quad \therefore \mathbb{E}_{\bar{\tau}^0, i, \epsilon} \left[\log \sigma \left(N \cdot C_i \left(\|\epsilon - \epsilon_{\theta}(\bar{\tau}^i, i)\|^2 - \|\epsilon - \epsilon_{\theta^{\text{old}}}(\bar{\tau}^i, i)\|^2 \right) \right) \right].
\end{aligned}$$

829 $C_i = \frac{\beta_i}{(1-\beta_i)(1-\alpha_i)}$, $\epsilon \sim \mathcal{N}(0, I)$ and $\bar{\tau}^i \sim q(\bar{\tau}^i | \bar{\tau}^0)$, thus, $\bar{\tau}^i = \sqrt{\alpha_i} \bar{\tau}^0 + (1 - \alpha_i) \epsilon$.

830 A.2 Monotonic Improvement

831 In this section, we show that maximizing the surrogate objective guarantees a monotonic improvement
832 in the original training objective. Revisiting Eq. (8), we denote the lower bound by $g_{\theta_k, \theta}(\tau^0)$, where
833 $\theta_k = \theta^{\text{old}}$:

$$\begin{aligned}
f_{\theta}(\tau^0) &= \log \sigma \left(\mathbb{E}_{\tau^{1:N} \sim q(\tau^{1:N} | \tau^0)} \log \frac{\prod_{i=1}^N p_{\theta}(\tau^{i-1} | \tau^i)}{\prod_{i=1}^N p_{\theta_k}(\tau^{i-1} | \tau^i)} \right) \\
&\geq \mathbb{E}_{i \sim \mathcal{U}(1, N), \tau^i \sim q(\tau^i | \tau^0)} \log \sigma \left(N \mathbb{E}_{\tau^{i-1} \sim q(\tau^{i-1} | \tau^i, \tau^0)} \log \frac{p_{\theta}(\tau^{i-1} | \tau^i)}{p_{\theta_k}(\tau^{i-1} | \tau^i)} \right) = g_{\theta_k, \theta}(\tau^0).
\end{aligned}$$

834 Since $\log \sigma(\cdot)$ is a concave function, the lower bound $g_{\theta_k, \theta}(\tau^0)$ is always less than or equal to the
835 original $f_{\theta}(\tau^0)$ for all $\theta \in \Theta$: (1) $f_{\theta}(\tau^0) \geq g_{\theta_k, \theta}(\tau^0)$. Moreover, when $\theta = \theta_k$, the log term in both
836 sides becomes zero, yielding (2) $f_{\theta_k}(\tau^0) = g_{\theta_k, \theta_k}(\tau^0) = \log 1/2$.

837 If we maximize $g_{\theta_k, \theta}(\tau^0)$ instead of $f_{\theta}(\tau^0)$:

$$\theta_{k+1} = \arg \max_{\theta \in \Theta} g_{\theta_k, \theta}(\tau^0),$$

838 then the following inequality holds:

$$f_{\theta_{k+1}}(\tau^0) \geq g_{\theta_k, \theta_{k+1}}(\tau^0) \geq g_{\theta_k, \theta_k}(\tau^0) = f_{\theta_k}(\tau^0).$$

839 Therefore, this procedure guarantees monotonic improvement of the original objective.

840 A.3 Comparison with DPO-Diffusion

841 DPO-Diffusion [32] is an algorithm for aligning diffusion models with human preferences by
842 considering ranked pairs (τ_w^0, τ_l^0) that indicate a preference for τ_w^0 over τ_l^0 . Built upon the Bradley-
843 Terry (BT) model and the bijectivity between reward and policy, DPO-Diffusion optimizes the
844 diffusion model p_{θ} with a reference distribution p_{ref} via:

$$\mathbb{E}_{(\tau_w^0, \tau_l^0)} \log \sigma \left(\mathbb{E}_{(\tau_w^{1:N}, \tau_l^{1:N})} \left[\log \frac{p_{\theta}(\tau_w^{1:N})}{p_{\text{ref}}(\tau_w^{1:N})} - \log \frac{p_{\theta}(\tau_l^{1:N})}{p_{\text{ref}}(\tau_l^{1:N})} \right] \right), \quad (14)$$

where $\tau_w^{1:N} \sim q(\cdot|\tau_w^0)$ and $\tau_l^{1:N} \sim q(\cdot|\tau_l^0)$. This objective function can be reformulated in terms of noise prediction as:

$$\mathbb{E}_{(\tau_w^0, \tau_l^0, i)} \log \sigma \left(N(\|\epsilon_w - \epsilon_{\text{ref}}(\tau_w^i, i)\|^2 - \|\epsilon_w - \epsilon_\theta(\tau_w^i, i)\|^2 + \|\epsilon_l - \epsilon_\theta(\tau_l^i, i)\|^2 - \|\epsilon_l - \epsilon_{\text{ref}}(\tau_l^i, i)\|^2) \right), \quad (15)$$

where ϵ_w and ϵ_l correspond to τ_w^i and τ_l^i , ϵ_θ and ϵ_{ref} are the noise prediction network for p_θ and p_{ref} .

By comparing the training objective functions, we can draw an interesting observation: if we denote expert samples as τ_w and generative samples as τ_l , with the reference model corresponding to the generator, the sigmoid in Eq. (15) is applied to the sum of the error difference on expert and generative samples. In contrast, DPAIL evaluates these two error differences in individual sigmoid functions.

This distinction arises from DPO-Diffusion’s derivation via the BT model, whereas DPAIL is based on the binary discriminator. Moreover, although DPO-Diffusion also aims to handle multi-modal distributions, it targets offline RL and requires preference data, setting it apart from DPAIL’s focus on direct expert imitation without preference annotations.

B Action Execution and Diffusion Sampling in DPAIL

In DPAIL, the diffusion policy generates fixed-horizon sub-trajectories of length H . The resulting H actions are executed sequentially in the environment, so action sequence generation is performed once every H environment steps. The action execution procedure is detailed in Algorithm 2. To condition on the current state at the start of the sampling process, we overwrite the corresponding state variable at each diffusion step with the current observed state. The sampling procedure is detailed in Algorithm 3.

Algorithm 2 Action execution

```

1:  $s_0 = \text{env.reset}()$ 
2: for step  $t = 0, 1, 2 \dots$  do
3:   if  $t \% H == 0$  then
4:     Sample actions  $a_{0:H} \sim p_{\theta^{\text{old}}}(\cdot|s_t)$ 
5:   end if
6:    $a_t \leftarrow \text{Get}(t \% H)\text{-th action in } a_{0:H}.$ 
7:    $r_t, s_{t+1} \leftarrow \text{env.step}(a_t)$ 
8: end for
```

Algorithm 3 Sampling

```

1: Observe the current state  $s_t, \tau^N \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ .
2: for diffusion step  $i = N, \dots, 1$  do
3:    $z \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 
4:    $\tau^{i-1} = \frac{1}{\sqrt{1-\beta_i}} \left( \tau^i - \frac{\beta_i}{\sqrt{1-\alpha_i}} \epsilon_{\theta^{\text{old}}}(\tau^i, i) \right) + \sigma_i z$ 
5:   Replace the initial state in  $\tau^i$  with  $s_t$ .
6: end for
7: Get action sequence  $a_{0:H}$  from  $\tau^0$ .
```

862

C Environment Details

HalfCheetah-v3 and Walker2d-v3 The goal of these tasks is to move the agent forward and backward along the x-axis as quickly as possible while maintaining balance. The state includes joint angles, angular velocities and the x-coordinate. Each expert is trained using a reward function based on the the forward reward, $\pm x$ -coordinate velocity.

Ant-v3 and AntGoal-v3 The goal of these tasks is to control a four-legged ant robot to move forward, backward, left, or right as fast as possible while maintaining balance (Ant-v3), and to navigate to one of eight target positions evenly distributed around a circle with a radius of 20 (AntGoal-v3). The state includes joint angles, angular velocities, and the (x,y) coordinates. Each expert is trained using a reward function based on the the forward reward, $\pm x$ -coordinate velocity or $\pm y$ -coordinate velocity in Ant-v3. In AntGoal-v3, each expert is trained using a reward function based on the distance between the goal and the current robot’s position. We use the 10 trajectories per mode as expert demonstrations, with each trajectory consisting of 1k transitions.

maze2d-medium-v1 and maze2d-large-v1 The goal of these tasks is to control a point robot to navigate to one of the target positions. In the medium-sized maze, target positions are $\{(1.0, 6.0), (6.0, 5.0), (6.0, 1.0)\}$, while in the large-sized maze, they are $\{(1.0, 10.0), (3.0, 8.0), (7.0, 10.0), (5.0,$

879 4.0), (7.0, 1.0)}. Expert demonstrations are selected from D4RL dataset ¹. We use the 15 episodes for
 880 maze2d-medium-v1 and 30 episodes for maze2d-large-v1.

881 D Implementation Details

882 **Policy gradient method** We use PPO [26] to train policies and GAE(λ) to compute advantage in
 883 GAIL, DiffAIL, DRAIL and InfoGAIL. The corresponding hyperparameters for PPO are provided in
 884 Table 3. At each k -th iteration, we perform m -steps rollout in the environment. The corresponding
 885 hyperparameter settings for each algorithm are provided in Table 2.

886 **Diffusion and DPAIL** Both Diffusion and DPAIL utilize the same U-Net architecture with residual
 887 blocks consisting of temporal convolution and group normalization, following [17] ². We use $N = 50$
 888 diffusion steps in both Diffusion and DPAIL for all tasks. Additionally, we normalize the state values
 889 before feeding them into the network. For DPAIL, we clip the norm value of $\|\epsilon - \epsilon_{\theta^{\text{old}}}(\bar{x}^i, i)\|$ not to
 890 be larger than 0.2.

891 **GAIL, DiffAIL, DRAIL and ASAF** For GAIL, DiffAIL, DRAIL, and ASAF, we use a multi-layer
 892 perceptron (MLP) with two hidden layers of size [64, 64] for the Gaussian policy. We also normalize
 893 the state values before feeding them into the policy network. The discriminator in GAIL is an MLP
 894 with two hidden layers of size [100, 100]. The discriminator architectures of both DiffAIL and
 895 DRAIL are based on an MLP U-Net structure based on the official repository ³, and $N = 50$ diffusion
 896 steps.

897 **InfoGAIL** For InfoGAIL, we use discrete latent variables, setting the number of latent variables
 898 to 8 for all tasks. We concatenate the one-hot encoding of the latent variable with the state and use
 899 the resulting vector as input to a Gaussian policy. We also normalize the state values before feeding
 900 them into the policy network. The discriminator network and class prediction network in InfoGAIL
 901 share an MLP with two hidden layers of size [100, 100] and output the corresponding values. For the
 902 coefficient of unsupervised regularization term, we perform a greedy search over the range [0.1, 0.2,
 903 0.3, 0.5].

904 **Form of the reward in GAIL, DiffAIL, DRAIL and InfoGAIL** For Mujoco tasks, we use a
 905 commonly adopted reward function of the form $r(s, a) = -\log(1 - D(s, a))$, which acts as a
 906 survival bonus, encouraging agents to survive longer in the environment to accumulate more rewards.
 907 For Maze tasks, we use the reward function $r(s, a) = \log(D(s, a))$, which serves as a penalty signal.
 908 This is well-suited for goal-reaching tasks, as it incentivizes the agent to reach the goal as quickly as
 909 possible. In AntGoal-v3, we adopt the survival-style reward $r(s, a) = -\log(1 - D(s, a))$ and we
 910 find this to work well in practice.

911 **Details of ASAF** ASAF aims to match the trajectory distribution under a stationary
 912 policy $\pi(a|s)$. For π , the trajectory distribution $p_\pi(\tau)$ is decomposed as $p_\pi(\tau) =$
 913 $P(s_0) \prod_{t=0}^{T-1} \pi(a_t|s_t) P(s_{t+1}|s_t, a_t)$. To optimize Eq. (4) for trainable policy π_θ and generator
 914 policy π_G , ASAF defines the discriminator in policy space as

$$D_{\pi_{\theta^{\text{old}}}, \pi_\theta}(\tau) = \sigma \left(\log \frac{p_{\pi_\theta}(\tau)}{p_{\pi_{\theta^{\text{old}}}}(\tau)} \right) = \sigma \left(\sum_t \log \pi_\theta(a_t|s_t) - \log \pi_{\theta^{\text{old}}}(a_t|s_t) \right) \quad (16)$$

915 where the transition probability $P(s_{t+1}|s_t, a_t)$ cancels out, leaving only to the ratio of policy terms.
 916 In practice, ASAF segments trajectories into windows of length w , updates π_θ via a binary cross-
 917 entropy loss, and then sets $\pi_{\theta^{\text{old}}}$ as the updated π_θ to the next iteration. This procedure iteratively
 918 updates π_θ until convergence. We offer the ASAF algorithm in Algorithm 4.

¹<https://github.com/Farama-Foundation/D4RL>

²<https://github.com/janner/diffuser>

³<https://github.com/NVlabs/DRAIL>

Algorithm 4 Adversarial Soft Advantage Fitting (ASAF)

Input: expert trajectories $\mathcal{D}_E = \{\tau_n\}_{n=1}^{N_E}$
Randomly initialize π_{θ_0} and set $\pi_{\theta^{\text{old}}} \leftarrow \pi_{\theta_0}$

for $k = \{0 \dots K\}$ **do**

Collect trajectories $\mathcal{D}_{\theta^{\text{old}}} = \{\bar{\tau}_n\}_{n=1}^{N_{\theta^{\text{old}}}}$ using $\pi_{\theta^{\text{old}}}$ by interacting with environment

Update θ_{k+1} by optimizing the following loss:

$$\theta_{k+1} = \arg \max_{\theta} \mathbb{E}_{\tau \sim \mathcal{D}_E} [\log D_{\pi_{\theta^{\text{old}}}, \pi_{\theta}}(\tau)] + \mathbb{E}_{\bar{\tau} \sim \mathcal{D}_{\theta^{\text{old}}}} [\log (1 - D_{\pi_{\theta^{\text{old}}}, \pi_{\theta}}(\bar{\tau}))],$$

where $D_{\pi_{\theta}, \pi_{\theta^{\text{old}}}}(\tau)$ is defined in Eq. (16).

$p_{\theta^{\text{old}}} \leftarrow p_{\theta_{k+1}}$

end for

Table 2: Hyperparameters used for baselines across various environments.

Method	Hyperparameter	HalfCheetah-v3	Walker2d-v3	Ant-v3	AntGoal-v3	maze2d-medium-v1	maze2d-large-v1
Diffusion	lr	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
	horizon H	4	4	4	4	16	16
	# Epoch	1000	1000	1000	1000	1000	1000
GAIL	policy lr	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003
	discriminator lr	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003
	# rollout length m	50000	50000	50000	50000	10000	10000
	# Iteration K	200	200	10000	600	100	100
DiffAIL	policy lr	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003
	discriminator lr	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
	# rollout length m	50000	50000	50000	50000	10000	10000
	# Iteration K	200	200	10000	600	100	100
DRAIL	policy lr	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003
	discriminator lr	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003
	# rollout length m	50000	50000	50000	50000	10000	10000
	# Iteration K	200	200	10000	600	100	100
InfoGAIL	policy lr	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003
	discriminator lr	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003
	coef MI	0.2	0.2	0.1	0.1	0.3	0.3
	# rollout length m	50000	50000	50000	50000	10000	10000
ASAF	# Iteration K	200	200	10000	600	100	100
	policy lr	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003
	window w	64	64	64	64	64	64
	# rollout length m	50000	50000	50000	50000	10000	10000
DPAIL (Ours)	# Iteration K	200	200	400	400	100	100
	lr	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
	horizon H	4	4	4	4	16	16
	# rollout length m	10000	10000	10000	10000	5000	5000
	# Iteration K	200	200	200	200	200	200

Table 3: PPO training hyperparameters used for each task.

Hyperparameter	HalfCheetah-v3	Walker2d-v3	Ant-v3	AntGoal-v3	maze2d-medium-v1	maze2d-large-v1
clipping range ϵ	0.2	0.2	0.2	0.2	0.2	0.2
discount factor γ	0.99	0.99	0.99	0.99	0.995	0.995
gae parameter λ	0.97	0.97	0.97	0.97	0.97	0.97
# epoch per iteration	50	50	50	50	40	40

919 E Additional Experimental Results

920 We present the learned behaviors of baseline methods and DPAIL in Figures 6 and 7. The expert
 921 demonstration behaviors are visualized in Figure 5. To evaluate the multi-modal learning capability
 922 of imitation learning methods, we measure the entropy of the learned behaviors to quantify diversity.
 923 Additionally, to assess the similarity between the learned trajectories and expert demonstrations, we
 924 compute Maximum Mean Discrepancy (MMD) between their respective state-action distributions,
 925 using an RBF kernel with 20 bandwidths, as shown in Table 4. Since MMD quantifies the divergence
 926 between distributions, lower values indicate better recovery of all modes present in the expert
 927 distribution.

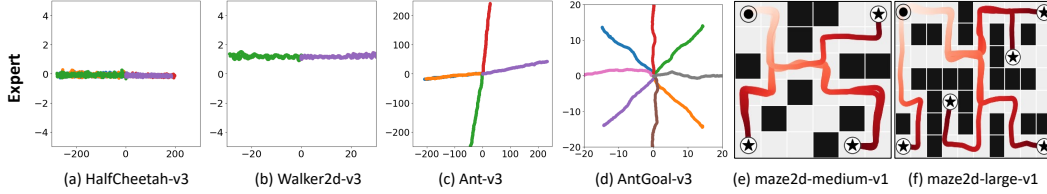


Figure 5: Expert demonstrations across 6 tasks.

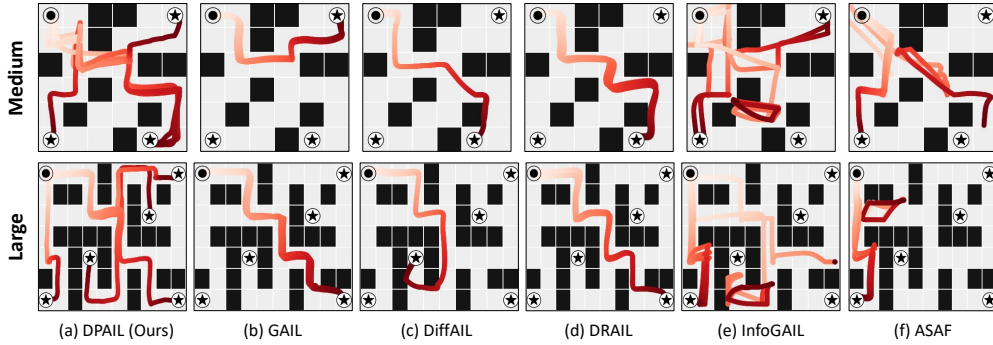


Figure 6: Learned behaviors of baseline methods and DPAIL (ours) in Maze2d tasks. The first row depicts maze2d-medium-v1, while the second row depicts maze2d-large-v1. Each graph illustrates 5 different trajectories generated by the same policy. The initial position is marked with circle, and the goal positions are marked with stars.

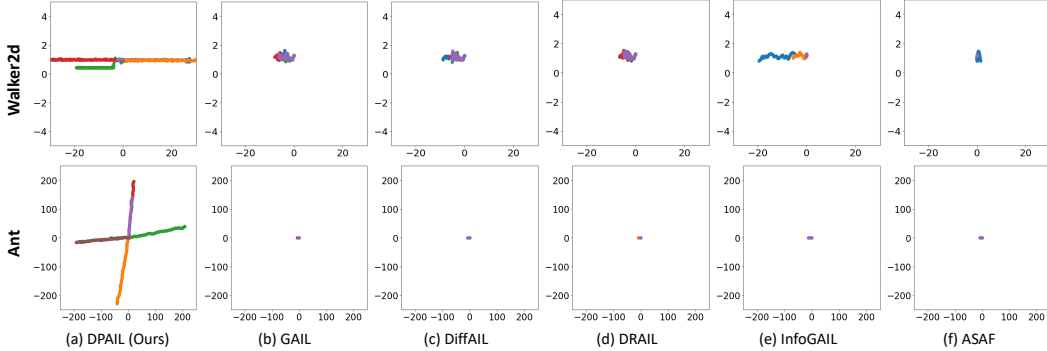


Figure 7: Learned behaviors of baseline methods and DPAIL (ours) in MuJoCo tasks. The first row shows Walker2d-v3, where the task is to move forward and backward ($\pm x$ -axis). The second row shows Ant-v3, where the task is to move forward, backward, left and right ($\pm x$ -axis, $\pm y$ -axis). Each plot illustrates ten different trajectories generated by the same policy.

928 E.1 Impact of the Number of Demonstrations and Modes

929 We provide additional experimental results on the impact of the number of demonstrations and
 930 modes in Ant-v3 (Table 5) and AntGoal-v3 (Table 6). In both tasks, increasing the number of modes
 931 generally degrades the performance of most algorithms. However, DPAIL exhibits greater robustness,
 932 benefiting from the expressiveness of diffusion models.

Table 4: MMD(\downarrow) between state-action distributions between expert demonstrations and learned behaviors. DPAIL has the lowest value on most tasks, indicating better recovery of expert distributions.

Environment	BC	Diffusion	GAIL	DiffAIL	DRAIL	InfoGAIL	ASAF	DPAIL
HalfCheetah-v3	0.029	0.027	0.038	0.038	0.045	0.039	0.031	0.025
Walker2d-v3	0.127	0.091	0.165	0.182	0.212	0.099	0.190	0.096
Ant-v3	0.334	0.018	0.281	0.341	0.335	0.282	0.311	0.012
AntGoal-v3	0.571	0.082	0.385	0.443	0.401	0.192	0.551	0.021
maze2d-medium-v1	4.9e-4	9.1e-5	4.8e-4	4.2e-4	4.5e-4	5.0e-4	5.5e-4	8.0e-5
maze2d-large-v1	5.1e-3	1.0e-4	3.0e-4	1.5e-4	5.0e-4	5.3e-4	5.9e-3	9.2e-5

Table 5: Normalized score (Score) and entropy (Ent) on varying the number of demonstrations and modes in Ant-v3.

# of modes	# of demos	metrics	BC	Diffusion	GAIL	DiffAIL	DRAIL	InfoGAIL	ASAF	DPAIL
1	3	Score	0.13 \pm 0.03	0.38 \pm 0.12	0.39 \pm 0.08	0.50\pm0.20	0.27 \pm 0.12	0.23 \pm 0.03	0.27 \pm 0.12	0.48 \pm 0.25
		Ent	0.00 \pm 0.00	0.00 \pm 0.00	0.00 \pm 0.00	0.00 \pm 0.00	0.00 \pm 0.00	0.00 \pm 0.00	0.00 \pm 0.00	0.00 \pm 0.00
	5	Score	0.15 \pm 0.06	0.4 \pm 0.23	0.39 \pm 0.08	0.55\pm0.25	0.22 \pm 0.15	0.21 \pm 0.01	0.22 \pm 0.15	0.53 \pm 0.22
		Ent	0.00 \pm 0.00	0.00 \pm 0.00	0.00 \pm 0.00	0.00 \pm 0.00	0.00 \pm 0.00	0.00 \pm 0.00	0.00 \pm 0.00	0.00 \pm 0.00
	10	Score	0.17 \pm 0.09	0.57 \pm 0.21	0.50 \pm 0.18	0.55 \pm 0.40	0.47 \pm 0.28	0.23 \pm 0.00	0.47 \pm 0.28	0.67\pm0.29
		Ent	0.00 \pm 0.00	0.00 \pm 0.00	0.00 \pm 0.00	0.00 \pm 0.00	0.00 \pm 0.00	0.00 \pm 0.00	0.00 \pm 0.00	0.00 \pm 0.00
	20	Score	0.10 \pm 0.09	0.74 \pm 0.08	0.51 \pm 0.12	0.69 \pm 0.14	0.45 \pm 0.24	0.25 \pm 0.07	0.45 \pm 0.24	0.72\pm0.05
		Ent	0.00 \pm 0.00	0.00 \pm 0.00	0.00 \pm 0.00	0.00 \pm 0.00	0.00 \pm 0.00	0.00 \pm 0.00	0.00 \pm 0.00	0.00 \pm 0.00
2	3	Score	0.10 \pm 0.06	0.40 \pm 0.22	0.17 \pm 0.19	0.14 \pm 0.29	0.28 \pm 0.10	0.12 \pm 0.01	0.10 \pm 0.12	0.47\pm0.10
		Ent	0.75 \pm 0.30	0.60 \pm 0.08	0.36 \pm 0.20	0.48 \pm 0.18	0.29 \pm 0.18	0.52 \pm 0.22	0.42 \pm 0.29	0.62 \pm 0.11
	5	Score	0.14 \pm 0.06	0.43 \pm 0.15	0.17 \pm 0.25	0.13 \pm 0.25	0.26 \pm 0.10	0.15 \pm 0.01	0.20 \pm 0.14	0.48\pm0.08
		Ent	0.88 \pm 0.09	0.49 \pm 0.24	0.34 \pm 0.18	0.46 \pm 0.26	0.31 \pm 0.18	0.53 \pm 0.26	0.46 \pm 0.32	0.60 \pm 0.13
	10	Score	0.18 \pm 0.15	0.52 \pm 0.09	0.15 \pm 0.23	0.26 \pm 0.29	0.31 \pm 0.23	0.13 \pm 0.02	0.15 \pm 0.02	0.64\pm0.08
		Ent	0.71 \pm 0.41	0.59 \pm 0.09	0.18 \pm 0.15	0.52 \pm 0.30	0.19 \pm 0.15	0.65 \pm 0.15	0.39 \pm 0.25	0.61 \pm 0.04
	20	Score	0.15 \pm 0.03	0.59 \pm 0.09	0.16 \pm 0.02	0.32 \pm 0.25	0.28 \pm 0.11	0.12 \pm 0.01	0.12 \pm 0.09	0.63\pm0.18
		Ent	1.01 \pm 0.25	0.66 \pm 0.01	0.44 \pm 0.25	0.51 \pm 0.23	0.36 \pm 0.32	0.52 \pm 0.18	0.44 \pm 0.30	0.59 \pm 0.08
3	3	Score	0.05 \pm 0.01	0.34 \pm 0.11	0.13 \pm 0.20	0.12 \pm 0.01	0.08 \pm 0.08	0.03 \pm 0.01	0.03 \pm 0.01	0.53\pm0.07
		Ent	0.93 \pm 0.32	0.85 \pm 0.11	0.76 \pm 0.17	0.82 \pm 0.19	0.98 \pm 0.10	0.95 \pm 0.08	0.81 \pm 0.12	1.01 \pm 0.16
	5	Score	0.06 \pm 0.05	0.44 \pm 0.13	0.11 \pm 0.21	0.10 \pm 0.24	0.10 \pm 0.05	0.02 \pm 0.01	0.01 \pm 0.02	0.55\pm0.03
		Ent	0.90 \pm 0.46	0.81 \pm 0.16	0.73 \pm 0.41	0.68 \pm 0.17	0.57 \pm 0.30	0.93 \pm 0.12	0.79 \pm 0.10	0.62 \pm 0.18
	10	Score	0.06 \pm 0.01	0.43 \pm 0.17	0.12 \pm 0.11	0.13 \pm 0.13	0.07 \pm 0.05	0.02 \pm 0.01	0.01 \pm 0.01	0.55\pm0.13
		Ent	0.90 \pm 0.29	0.91 \pm 0.16	0.82 \pm 0.18	0.90 \pm 0.07	0.69 \pm 0.37	0.99 \pm 0.08	0.75 \pm 0.38	1.11 \pm 0.50
	20	Score	0.03 \pm 0.04	0.60 \pm 0.04	0.11 \pm 0.13	0.15 \pm 0.05	0.12 \pm 0.09	0.01 \pm 0.00	0.05 \pm 0.06	0.64\pm0.14
		Ent	1.00 \pm 0.24	0.88 \pm 0.06	0.83 \pm 0.14	0.96 \pm 0.12	0.98 \pm 0.09	1.06 \pm 0.03	0.65 \pm 0.48	0.91 \pm 0.14
4	3	Score	0.08 \pm 0.01	0.21 \pm 0.11	0.08 \pm 0.03	0.02 \pm 0.01	0.02 \pm 0.01	0.03 \pm 0.01	0.00 \pm 0.00	0.39\pm0.06
		Ent	0.90 \pm 0.27	1.18 \pm 0.05	0.85 \pm 0.42	1.22 \pm 0.08	1.06 \pm 0.05	1.07 \pm 0.08	1.01 \pm 0.19	1.20 \pm 0.11
	5	Score	0.09 \pm 0.01	0.28 \pm 0.09	0.01 \pm 0.00	0.04 \pm 0.05	0.07 \pm 0.09	0.02 \pm 0.01	0.01 \pm 0.00	0.42\pm0.01
		Ent	0.95 \pm 0.21	1.17 \pm 0.06	0.95 \pm 0.09	1.31 \pm 0.04	1.07 \pm 0.70	1.19 \pm 0.08	0.89 \pm 0.16	1.20 \pm 0.14
	10	Score	0.06 \pm 0.07	0.48 \pm 0.11	0.01 \pm 0.00	0.02 \pm 0.01	0.02 \pm 0.08	0.03 \pm 0.02	0.00 \pm 0.00	0.56\pm0.06
		Ent	0.75 \pm 0.22	1.22 \pm 0.12	1.19 \pm 0.10	1.21 \pm 0.10	1.26 \pm 0.11	1.04 \pm 0.07	1.07 \pm 0.19	1.21 \pm 0.05
	20	Score	0.05 \pm 0.05	0.60 \pm 0.02	0.01 \pm 0.00	0.04 \pm 0.02	0.03 \pm 0.01	0.02 \pm 0.02	0.00 \pm 0.00	0.65\pm0.04
		Ent	1.01 \pm 0.36	1.13 \pm 0.08	1.29 \pm 0.03	1.31 \pm 0.06	1.30 \pm 0.05	1.31 \pm 0.05	0.87 \pm 0.35	1.18 \pm 0.15

Table 6: Normalized score (Score) and entropy (Ent) on varying the number of demonstrations and modes in AntGoal-v3.

# of modes	# of demos	metrics	BC	Diffusion	GAIL	DiffAIL	DRAIL	InfoGAIL	ASAF	DPAIL
1	3	Score	0.39 ± 0.17	0.62 ± 0.18	0.74 ± 0.08	0.94 ± 0.02	0.78 ± 0.10	0.62 ± 0.04	0.40 ± 0.21	0.80 ± 0.02
		Ent	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
	5	Score	0.43 ± 0.20	0.65 ± 0.21	0.84 ± 0.09	0.96 ± 0.02	0.83 ± 0.05	0.68 ± 0.09	0.41 ± 0.38	0.83 ± 0.08
		Ent	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
	10	Score	0.46 ± 0.14	0.79 ± 0.05	0.92 ± 0.02	0.94 ± 0.02	0.93 ± 0.04	0.64 ± 0.13	0.47 ± 0.26	0.82 ± 0.11
		Ent	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
	20	Score	0.52 ± 0.20	0.84 ± 0.04	0.86 ± 0.05	0.94 ± 0.01	0.80 ± 0.07	0.65 ± 0.00	0.53 ± 0.16	0.88 ± 0.01
		Ent	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
2	3	Score	0.08 ± 0.03	0.51 ± 0.08	0.68 ± 0.22	0.81 ± 0.04	0.72 ± 0.07	0.45 ± 0.06	0.02 ± 0.01	0.69 ± 0.10
		Ent	0.50 ± 0.13	0.46 ± 0.23	0.20 ± 0.17	0.54 ± 0.09	0.23 ± 0.10	0.54 ± 0.21	0.51 ± 0.27	0.68 ± 0.12
	5	Score	0.10 ± 0.08	0.54 ± 0.19	0.74 ± 0.11	0.86 ± 0.08	0.74 ± 0.13	0.47 ± 0.10	0.04 ± 0.04	0.71 ± 0.07
		Ent	0.56 ± 0.12	0.42 ± 0.33	0.15 ± 0.11	0.64 ± 0.05	0.20 ± 0.09	0.64 ± 0.17	0.53 ± 0.31	0.71 ± 0.17
	10	Score	0.13 ± 0.09	0.58 ± 0.21	0.77 ± 0.07	0.80 ± 0.04	0.72 ± 0.05	0.49 ± 0.12	0.03 ± 0.03	0.80 ± 0.06
		Ent	0.68 ± 0.10	0.42 ± 0.24	0.17 ± 0.17	0.53 ± 0.14	0.15 ± 0.13	0.68 ± 0.08	0.65 ± 0.25	0.52 ± 0.18
	20	Score	0.15 ± 0.09	0.83 ± 0.09	0.75 ± 0.11	0.86 ± 0.07	0.76 ± 0.13	0.49 ± 0.07	0.16 ± 0.15	0.87 ± 0.04
		Ent	0.54 ± 0.13	0.33 ± 0.18	0.43 ± 0.18	0.62 ± 0.09	0.43 ± 0.16	0.61 ± 0.16	0.51 ± 0.17	0.52 ± 0.15
4	3	Score	0.01 ± 0.00	0.37 ± 0.12	0.52 ± 0.13	0.52 ± 0.21	0.54 ± 0.22	0.40 ± 0.28	0.02 ± 0.01	0.64 ± 0.10
		Ent	0.94 ± 0.05	0.99 ± 0.13	1.02 ± 0.28	1.02 ± 0.09	1.04 ± 0.20	1.19 ± 0.16	0.92 ± 0.28	1.14 ± 0.26
	5	Score	0.03 ± 0.01	0.39 ± 0.26	0.54 ± 0.09	0.50 ± 0.19	0.50 ± 0.17	0.38 ± 0.11	0.03 ± 0.04	0.66 ± 0.07
		Ent	0.97 ± 0.03	1.00 ± 0.15	1.00 ± 0.38	1.09 ± 0.18	1.05 ± 0.18	1.23 ± 0.15	0.82 ± 0.30	1.16 ± 0.27
	10	Score	0.04 ± 0.02	0.45 ± 0.28	0.67 ± 0.95	0.66 ± 0.08	0.67 ± 0.11	0.52 ± 0.14	0.15 ± 0.31	0.75 ± 0.06
		Ent	1.02 ± 0.07	1.05 ± 0.06	0.88 ± 0.21	0.94 ± 0.24	0.96 ± 0.25	1.25 ± 0.12	0.56 ± 0.42	1.05 ± 0.17
	20	Score	0.02 ± 0.01	0.39 ± 0.26	0.54 ± 0.09	0.50 ± 0.19	0.50 ± 0.17	0.38 ± 0.11	0.03 ± 0.04	0.66 ± 0.07
		Ent	0.98 ± 0.26	1.00 ± 0.15	1.00 ± 0.38	1.09 ± 0.18	1.05 ± 0.18	1.23 ± 0.15	0.82 ± 0.30	1.16 ± 0.27
8	3	Score	0.02 ± 0.01	0.20 ± 0.07	0.43 ± 0.11	0.32 ± 0.04	0.39 ± 0.07	0.39 ± 0.11	0.01 ± 0.00	0.52 ± 0.05
		Ent	1.26 ± 0.10	1.50 ± 0.38	1.48 ± 0.27	1.78 ± 0.12	1.79 ± 0.23	1.77 ± 0.05	1.43 ± 0.33	1.70 ± 0.27
	5	Score	0.03 ± 0.01	0.23 ± 0.09	0.47 ± 0.06	0.37 ± 0.08	0.40 ± 0.10	0.45 ± 0.09	0.01 ± 0.00	0.54 ± 0.03
		Ent	0.98 ± 0.20	1.56 ± 0.48	1.78 ± 0.17	1.82 ± 0.13	1.81 ± 0.13	1.79 ± 0.08	1.65 ± 0.23	1.76 ± 0.41
	10	Score	0.04 ± 0.04	0.22 ± 0.07	0.58 ± 0.05	0.35 ± 0.16	0.41 ± 0.15	0.48 ± 0.07	0.01 ± 0.00	0.67 ± 0.03
		Ent	1.46 ± 0.58	1.52 ± 0.31	1.51 ± 0.17	1.75 ± 0.11	1.72 ± 0.07	1.78 ± 0.13	1.52 ± 0.27	1.73 ± 0.23
	20	Score	0.01 ± 0.00	0.53 ± 0.18	0.58 ± 0.02	0.45 ± 0.08	0.41 ± 0.10	0.46 ± 0.02	0.01 ± 0.00	0.74 ± 0.02
		Ent	1.51 ± 0.24	1.18 ± 0.41	1.30 ± 0.14	1.75 ± 0.23	1.79 ± 0.12	1.79 ± 0.12	1.46 ± 0.16	1.78 ± 0.41