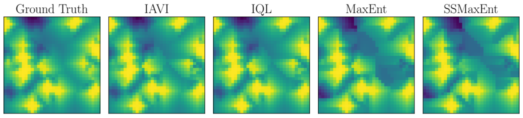
We thank all reviewers for their very constructive and valuable feedback and their insightful comments.

Reviewer 1: The supplementary material does not contain the source code. Due to confidentiality, we had to wait for permission to release code, which we got in the meantime. **We will release source code upon camera ready.**

Reviewer 2, 3 and 4: One weakness of the approach is the underlying assumption that the expert demonstrations are generated by a Boltzmann distribution. Most likely Eq. 1 in the paper led to some confusion here. To clarify: the expert demonstrations can follow an **arbitrary** distribution (including multimodal ones and greedy behavior choice), and need not necessarily follow a Boltzmann distribution. Our approach encodes this underlying arbitrary distribution of the expert in the long-term Q-value, such that a Boltzmann distribution over the estimated Q-function is equivalent to the original arbitrary expert distribution. **Put differently, our assumption is a restriction over the space of Q-functions, not expert policies** (see proof below). **We will clarify this and add the proof in the camera-ready.**

Theorem 1. *Define $Q^*(s,a)$ to satisfy the equation $Q^*(s,a) = Q^*(s,b) + \log(\pi^{\mathcal{E}}(a|s)) - \log(\pi^{\mathcal{E}}(b|s))$ for all actions $a, b \in \mathcal{A}$ (as in Section 3 ff.) and expert policy $\pi^{\mathcal{E}}(\cdot|s)$ of arbitrary underlying distribution. Then the Boltzmann distribution over $Q^*(s,\cdot)$ is equivalent to $\pi^{\mathcal{E}}(\cdot|s)$.* ***Proof.*** *The theorem follows from the inverse application of Equations (1)-(3). With $Q^*(s,a) = Q^*(s,b) + \log(\pi^{\mathcal{E}}(a|s)) - \log(\pi^{\mathcal{E}}(b|s))$ for all actions $a, b \in \mathcal{A}$, it follows from Eq. (3) that* $\exp\left(Q^*(s,a)\right) = \left(\pi^{\mathcal{E}}(a|s)/\pi^{\mathcal{E}}(b|s)\right)\exp\left(Q^*(s,b)\right) = \pi^{\mathcal{E}}(a|s)\sum_{A\in\mathcal{A}}\exp(Q^*(s,A))$ *and thus* $\pi^{\mathcal{E}}(a|s) = \exp(Q^*(s,a))/\sum_{A\in\mathcal{A}}\exp(Q^*(s,A))$.

Reviewer 2 and 4: What about a greedy expert policy? We actually addressed the case of strictly optimal demonstrations in the SUMO experiments, Sect. 7.2. We further provide results for a greedy expert policy in Objectworld (Fig. 1 below). IAVI outperforms MaxEnt IRL also in this setting by multiple orders of magnitude w.r.t. runtime, and both IAVI and IQL yield a smaller EVD after less training time. **To be added to the camera-ready.**



	IAVI	IQL	MaxEnt	SSMaxEnt
EVD	**0.22**	**0.21**	2.8	2.96
Runtime	**0.03 h**	**0.71 h**	5.02 h	4.65 h

Figure 1: Results for a greedy expert policy in Objectworld. Visualization and table as in Figure 3 in the main paper.

Reviewer 3: What if there is a significant (suboptimal) bias in the expert policy? This is an open problem common to almost all published IRL methods which are based on the assumption of (soft-)optimal expert demonstrations, including e.g. GAIL or Maximum Entropy methods. We are aware that addressing systematic bias is important for future work, but out of scope for this work. Known bias could be accounted for by shifting the log probabilities of the original expert distribution. **We will add a comment to the camera-ready.**

Reviewer 2: Are IAVI and IQL guaranteed to converge? What theoretical guarantees can you provide? Do they reach policy matching? For expert policies with only non-zero action probabilities, we showed that the immediate reward function defined in Section 3 leads to a Boltzmann distribution over Q-values which reflects the expert policy. Since the Bellman update is a contraction, the reward values become more accurate with each iteration. Convergence guarantees for IAVI and IQL then follow from the convergence guarantees of Value Iteration and Q-learning under the same conditions. If there are actions with zero probability mass under the expert demonstrations, we add a very small conditioning term ϵ to the probability to avoid numerical instabilities. This leads us back to the case above with guaranteed convergence, but introduces a small deviation with respect to the match of the expert policy, bounded in dependence of ϵ. Empirically, we showed convergence in our experiments in Section 7.1 und 7.2. **We will add more details to the camera-ready.**

Reviewer 2 and 3: Can you extend IQL to Soft Q-learning or continuous action spaces? Initial experiments with entropy regularization led to minor improvements. Our method can also readily be extended to continuous action-spaces. We currently evaluate a first working version. However, we regard these extensions as out of scope for the current submission. Results will be presented in future publications. **We will add a comment to the camera-ready.**

Reviewer 2: Can't you calculate the action probabilities beforehand? We formalized the more general case of an online algorithm, i.e. the possibility to add transitions during training. We explored the offline case of estimating the visitation probabilities beforehand in initial experiments with good results. **We will add this to the appendix.**

Reviewer 4: The math is hard to follow. **We will try to revise the math to add clarity in the camera ready**.