We would like to thank all reviewers for their constructive feedback and insightful comments. Here our answers:

Reviewer #1 As you stated correctly, the main contribution of this paper lies in the generality of the proposed inference framework, not in achieving higher scores than domain-specific algorithms. Nevertheless, we agree that a comparison with such algorithms will lead to a much stronger conclusion (see also Reviewer 3) and we will try our best to include additional baselines in the final version. Furthermore, we will make our code available with the camera-ready version.

We are aware that there exist approaches (like IRL) that are capable of generalizing behavior to unobserved situations by shifting the reasoning to the intentional level. However, please note that the very same logic also applies on that level: explicitly taking into account correlations between intentions (e.g. given in form of subgoals) can lead to better predictive models that outperform their correlation-agnostic counterparts while requiring less training data. To demonstrate this effect, we will include an additional IRL scenario in the camera-ready version, which we had to omit in the submitted version for space reasons. For details, please see comments to Reviewer 2 and preliminary results in the figure below.

Reviewer #2 We agree that the range of modeling scenarios to which our methodology can be applied is quite open. However, we consider this as a clear benefit as it emphasizes the generality of our approach, which enables many possible future developments. While planar navigation on a grid is a nice toy problem to demonstrate the working principle, the same methodology can be applied to arbitrary decision-making problems comprising discrete-valued elements. Moreover, apart from single-agent RL, we have many other potential use cases in mind, such as modeling correlations between agents in a multi-agent network and learning common structures across tasks (transfer learning).

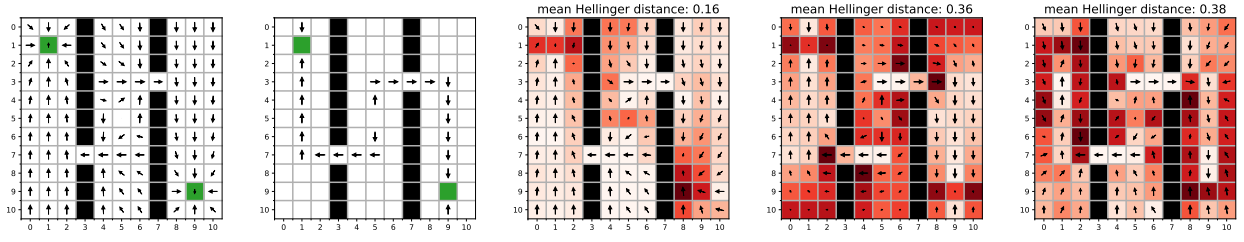
In general, the proposed methodology can be beneficial in all scenarios that contain some form of structure (a wall in a grid world is just one particular example, just like the underlying grid layout itself and the associated translation-invariant transition dynamics). Often, these structures manifest themselves at different levels (e.g. similar intentions triggering similar actions) and applying our concept on one particular level can be superior to others, depending on the application. This effect is demonstrated by the situation shown in the figure below (preliminary results), which compares the PG imitation learning model from Section 4.1 that captures action correlations with one where the PG concept has been applied on the intentional level, capturing correlations between subgoals assigned to different states. In fact, the latter describes an additional use case of our framework, which we had to exclude from the paper due to space limitations but which will be described in detail in the camera-ready version (which is possible due to one additional page of content). The example shows a clear improvement for the subgoal-based variant, which generalizes the data on the intentional level (and not on the action level), yielding a significantly better reconstruction in unobserved regions of the state space. Note: the Dirichlet imitation learning model is omitted as its prediction is restricted to the visited states (compare Fig. 1b in the paper).

Reviewer #3 Many thanks for your feedback. Here are a few things that we would like to mention:

• We are eager to apply our approach to larger domains and real data. However, this requires further model approximations, which are out of the scope of this paper and which we leave for future work. Due the involved covariance matrix inversion, inference in our model scales cubically with the cardinality C of the considered covariate space, currently preventing any large-scale experiments (similar to GP inference). Nevertheless, approximations similar to those used for GPs are possible, e.g. using local correlation models (compare inducing point method for GPs) or conjugate gradients.

• Please note that the derivations found in [12] (arXiv version) do neither include the variational principle for this particular problem type (instead, an LDA-like scenario is considered) nor the corresponding hyper-parameter optimization.

• We will discuss the suggested methods (Poupart and Texplore, see Reviewer 1) in the related work section of the camera-ready version and try to establish a performance comparison with our approach. However, note that a fair comparison might be difficult due to the conceptual differences of all approaches, i.e. the mentioned baseline do not capture the correlation structure explicitly and hyper-parameters need to be specified manually for each problem domain.



(a) expert policy & rewards (b) data set & rewards (c) policy reconstruction via **PG subgoal modeling** (d) policy reconstruction via **PG imitation learning** (e) policy reconstruction via **Dirichlet subgoal modeling**

Figure 1: Modeling action correlations vs. modeling subgoal correlations. Black squares indicate wall states. Red color indicates the per-state policy reconstruction in terms of the Hellinger distance to the expert policy.