A More results for Gotoobj

We use this part of the appendix to give additional figures for the Gotoobj level.

First, similarly to the reduction in policy loss as investigated in section 5, we are also interested in the reduction of overall uncertainty (Shannon entropy) over actions that the learner achieves by asking for guidance and compare this again for the cases where the learner actually wants to open or close the gate. This can be found in Figure 12. Note however the subtlety that the learner may become more certain about which actions to choose, but focus on the wrong action. By spotting differences between the two measures we may identify situations in which the guidance misleads the learner or, vice versa, in which a very certain learner that aims for the wrong action actually gets less certain by getting the correct guidance, leading to a reduction in policy loss.

As we see, the shape of the resulting plots are relatively similar to those of the policy loss. As one notable difference, the counterfactual entropy does not temporarily increase in the same way as the policy loss for open gate situations. This indicates that the learner generally gets continuously more certain, albeit not necessarily about the right actions.

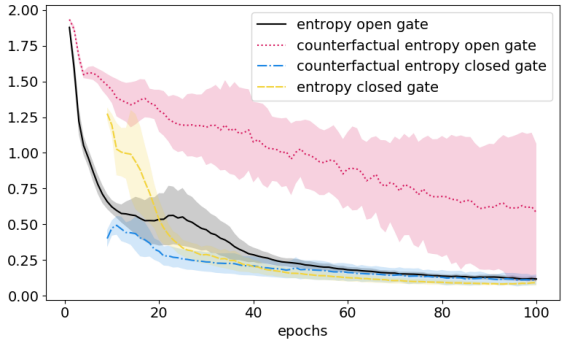


Figure 12: Entropy Comparison for GoToObj

Second, in figure 13 we give a more detailed view of the development of the spatial frequency of guidance requests as discussed in section 5.2.

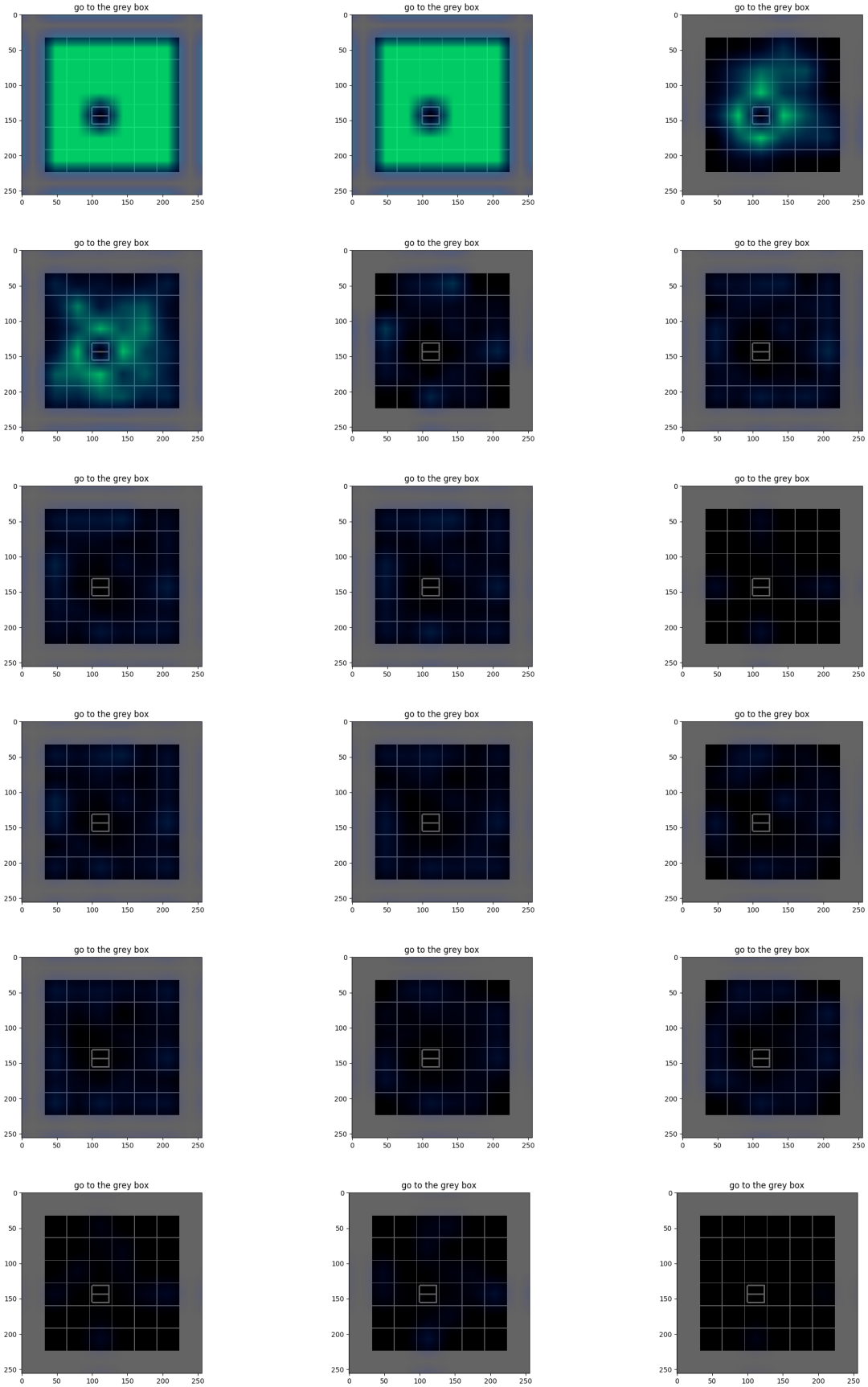
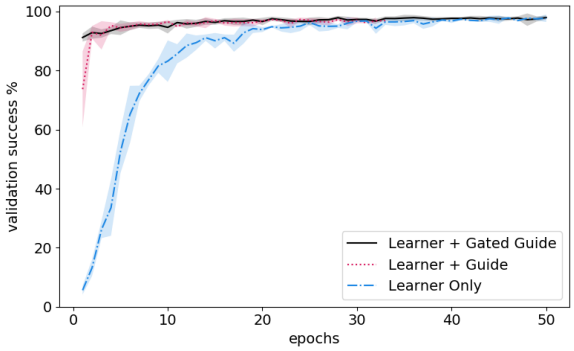


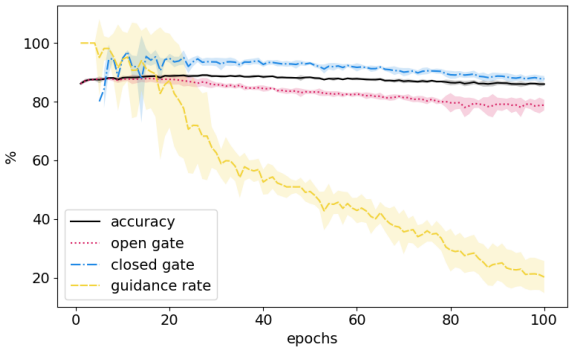
Figure 13: Heatmaps for GoToObj. They are ordered from left to right and then top to bottom. This shows how the guidance requests evolve over the course of the whole training in one specific example mission. Over time, not much guidance remains.

B Results for Putnextlocal

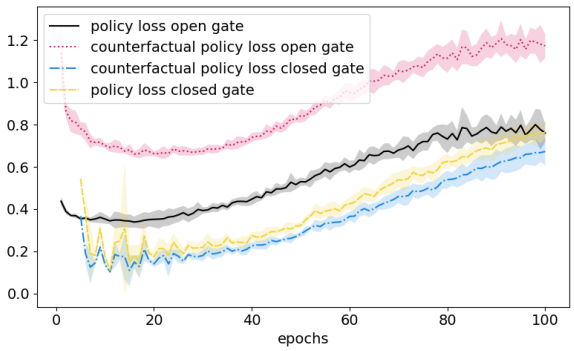
We provide results for PutNextLocal corresponding to the same results for GoToObj which were in the body of the paper. As is visible, many of our findings remain valid in this higher level. Note however that over the course of training, we observe significant overfitting. We nevertheless showed the whole development of the learning process in order to show the full development of the guidance rate.



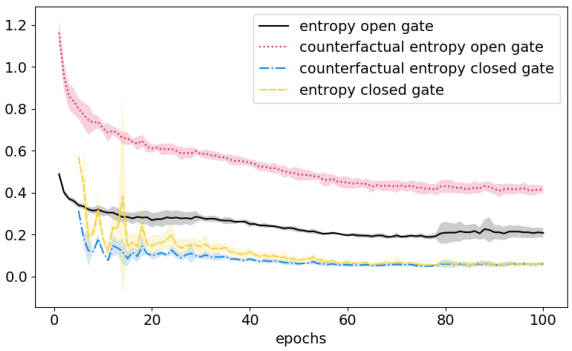
(a) Baseline comparison



(b) Validation Accuracy Comparison



(c) Policy Loss comparison



(d) Entropy comparison

Figure 14: PutNextLocal

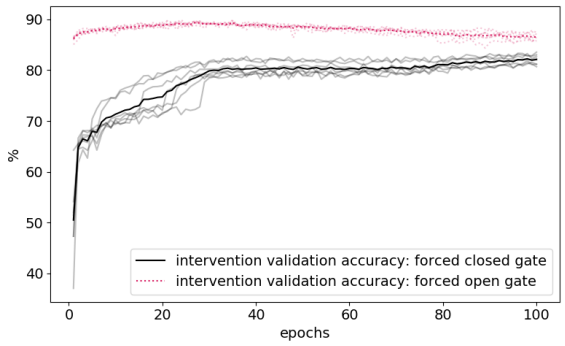


Figure 15: PutNextLocal: We compare the accuracy during validation in cases of forced open and closed gates: irrespective of the gating weight g_t computed from the system, we set $g_t = 1$ (so that the policy bases its decision on the encoded guidance $\text{Enc}(m_t)$) for the red dotted curve and $g_t = 0$ for the black curve.

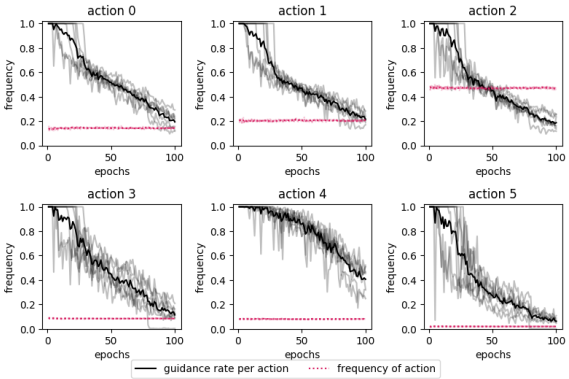


Figure 16: PutNextLocal: Frequency of open gate conditioned on actions and frequencies of actions themselves.

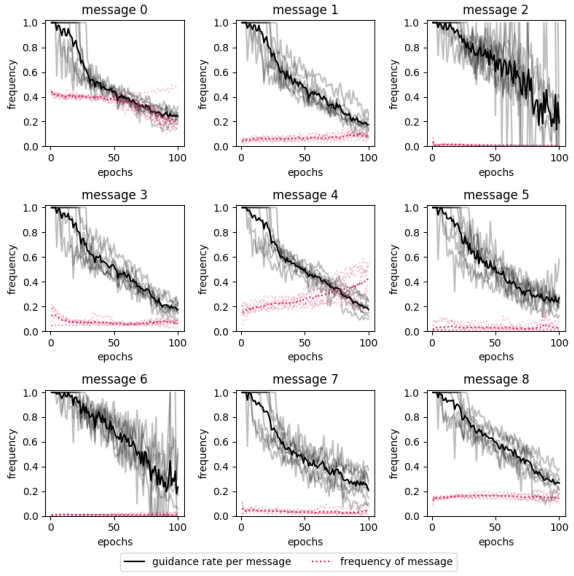


Figure 17: PutNextLocal: Frequency of open gate conditioned on messages and frequencies of messages themselves.

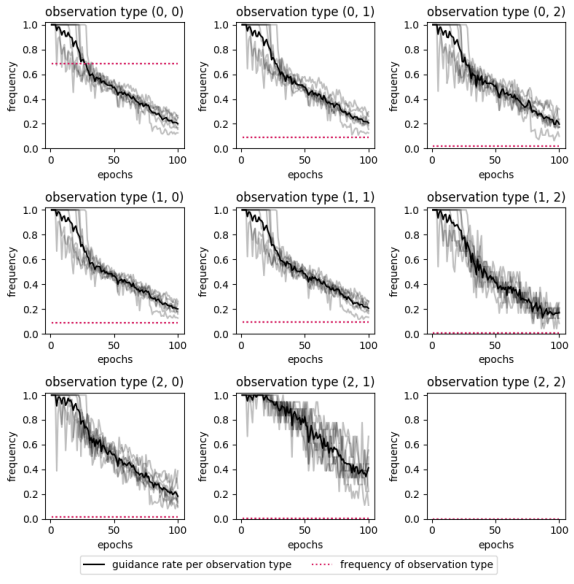


Figure 18: PutNextLocal: Frequency of open gate conditioned on observation types and frequencies of observation types themselves. For example, type $(2, 1)$ is a situation where directly left of the agent there is the goal and right of it there is an object sharing one feature with the goal-object.

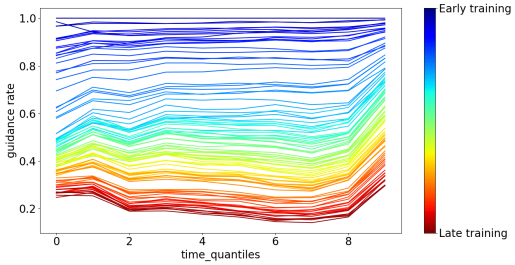


Figure 19: PutNextLocal: Guidance per time quantile: roughly speaking, a timepoint t is in quantile k of 10 if $t/l \approx k/10$, where l is the length of the corresponding episode. The plots show the guidance rate corresponding to the different quantiles. Dark blue curves belong to earlier epochs whereas red curves belong to later epochs.

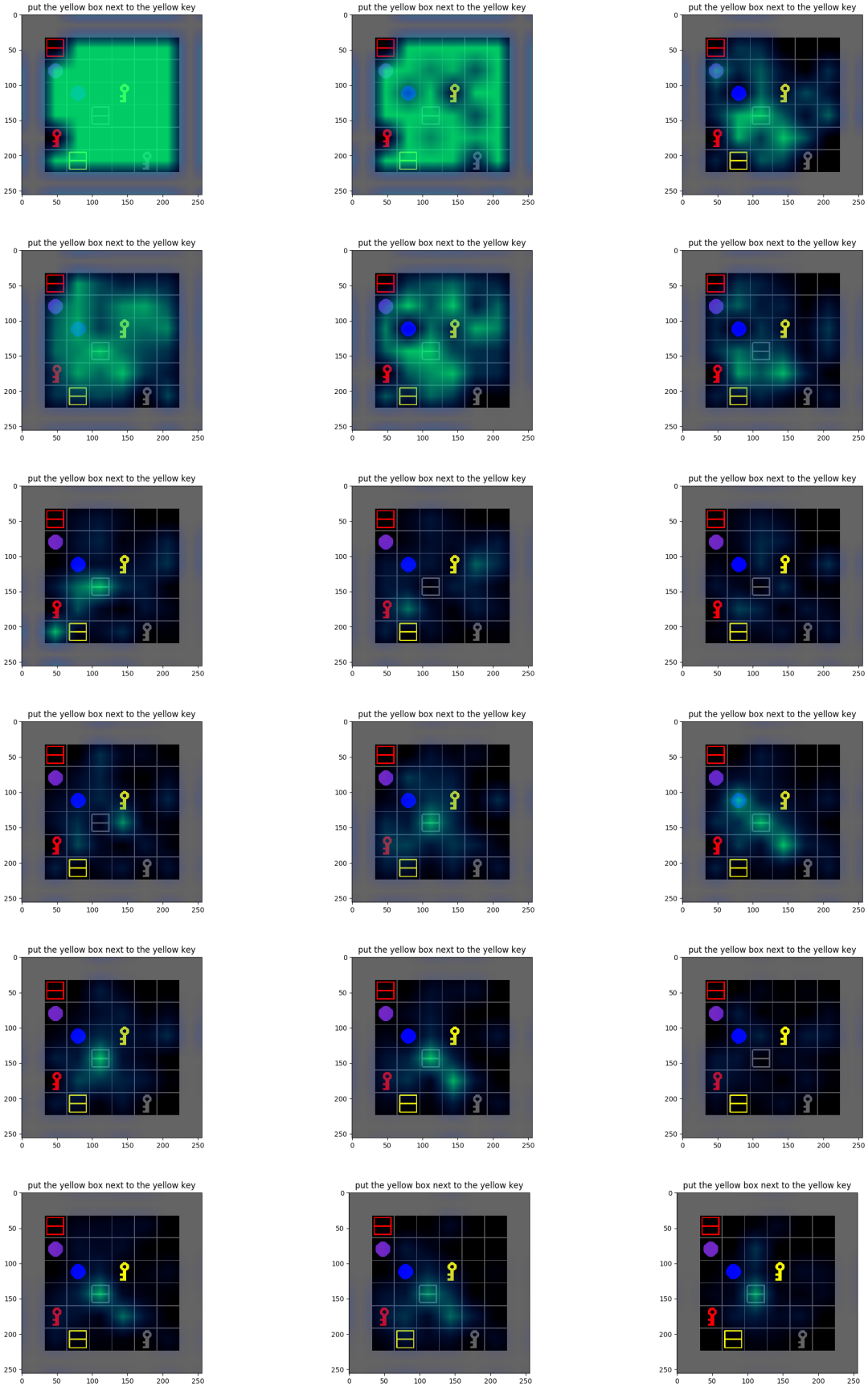


Figure 20: Heatmaps for PutNextLocal. They are ordered from left to right and then top to bottom. This shows how the guidance requests evolve over the course of the whole training in one specific example mission.