

Learning to Request Guidance in Emergent Communication

Benjamin Kolb,* Leon Lang*, Henning Bartsch, Arwin Gansekoele,
Raymond Koopmanschap, Leonardo Romor, David Speck, Mathijs Mul[†], Elia Bruni[†]

{benjamin.kolb, leon.lang, henning.bartsch,
arwin.gansekoele, raymond.koopmanschap, leonardo.romor,
david.speck, mathijs.mul}@student.uva.nl
elia.bruni@gmail.com

University of Amsterdam

Abstract

Previous research into agent communication has shown that a pre-trained guide can speed up the learning process of an imitation learning agent. The guide achieves this by providing the agent with discrete messages in an emerged language about how to solve the task. We extend this one-directional communication by a one-bit communication channel from the learner back to the guide: It is able to ask the guide for help, and we limit the guidance by penalizing the learner for these requests. During training, the agent learns to control this gate based on its current observation. We find that the amount of requested guidance decreases over time and guidance is requested in situations of high uncertainty. We investigate the agent’s performance in cases of open and closed gates and discuss potential motives for the observed gating behavior.

1 Introduction

A long-term goal of AI is to develop agents that can help humans execute complex tasks in the real world. Since reward functions that are aligned with human intentions are hard to manually specify (Amodei and Clark, 2016), other approaches besides Reinforcement Learning are needed for creating agents that behave in the intended way. Among these are Reward Modeling (Leike et al., 2018) and Imitation Learning (Pomerleau, 1991), but, eventually, it would be useful if we could use natural language to transmit wishes to the agents.

Recently, Mul et al. (2019) made progress in this direction by showing how communication can be used to guide a learner in a gridworld environment. Using emergent discrete messages, the guide is able to speed up the learning process of the learner and to let it generalize across incrementally more difficult environments.

In this setting, the communication channel is completely one-way: in each time step, the guide transmits a message that may help the learner make its decisions. In reality, however, communication is more complex than that: the guidance may be expensive and it can, therefore, be beneficial to have more sparse messages. Furthermore, the learner may want to ask for clarification if something is unclear. Therefore, it would be worthwhile if there was a communication channel from the learner back to the guide. It is this interactive nature of communication that arguably is needed for more advanced AI systems (Mikolov et al., 2016).

In this paper, we equip the learner introduced by Mul et al. (2019) with a binary gate to indicate its need for guidance in each time step. A penalty for the use of guidance incentivizes a sparse usage of the gate. By analyzing the relationship between the learner’s usage of the gate and a number of measures, we show that the learner indeed learns to ask for guidance in a smart and economical way.

2 Related Work

In this section we briefly lay out relevant work that relate to our approach on the dimensions of following language instructions, emergent communication and the interactions that emerge from guidance requests.

2.1 Following Language Instruction

In recent years, much research has been conducted in the field of following language instructions. Since language commands build a way to interact with agents and communicate information in an effective and human-interpretable way, the agent’s processing and understanding of these commands is relevant for our project. Starting with manually engineered mappings (Winograd,

*Equal contributions

[†]Shared senior authorship

1971), the currently most relevant grounded language acquisition methods focus on learning a parser to map linguistic input to its executable equivalent, i.e. action specifications using statistical models (Yu et al., 2018), (Kim and Mooney, 2012), (Artzi and Zettlemoyer, 2013), (Mei et al., 2016). In the BabyAI platform introduced by Chevalier-Boisvert et al. (2019) a synthetic “Baby Language” is used, which consists of a subset of English and whose semantics is generated by a context-free grammar (as opposed to instruction templates) and is easily understood by humans. They employ a single model to combine linguistic and visual information, similar to Misra et al. (2017). Our setup builds on Mul et al. (2019) who extend that platform with a guide, like Co-Reyes et al. (2018), that supplements the agent’s information with iterative linguistic messages.

2.2 Emergent Communication

In order to benefit most from the guide, the agent would ideally communicate back, thus creating a multi-agent cooperation scenario. Recent research in this area investigates the emergence and usage of emergent language, e.g. in the context of referential games (Lazaridou et al., 2016). Furthermore, Mordatch and Abbeel (2018) show that multiple agents can develop a grounded compositional language to fulfill their tasks more effectively with spoken exchange. In our setup the emergent communication consists of discrete word tokens similar to Havrylov and Titov (2017). Jiang and Lu (2018) propose an attentional communication model to learn when communication is needed (and helpful), resulting in more effective (large-scale) multi-agent cooperation.

2.3 Guidance Requests

In prior work (Mul et al., 2019), guidance is given at every time step and the communication is one-way from guide to learner. We extend this approach by allowing a communication channel in the other direction. Here we survey work that uses similar requests for help.

Most similar to our work is Clouse (1997), where “Ask for help” is proposed: in this setting, an RL agent has one additional action with which it can signify to a “trainer” that it wants to receive help. The trainer then chooses the agent’s action. Whether to ask for help is based on uncertainty about the highest action value. This is different from our setting in which the uncertainty

is only *implicitly* responsible for queries, as can be seen in Section 5. Kosoy (2019) studies the “Ask for help” setting theoretically and proves a regret bound for agents that act in infinite-horizon discounted MDPs and are able to delegate actions to an “advisor”.

In Nguyen et al. (2019) there is a help-requesting policy π_{help} that can signify if the agent needs help. If this is the case, a guide answers with a language-based instruction of subgoals. Additionally, there is a budget that limits asking for help.

Also related is Werling et al. (2015), where structured prediction problems are considered: a sequence of words is received and each word is supposed to be mapped to a label. The system can query a crowd (as in crowd-sourcing) to obtain answers on specific words in the sequence. As in our case, querying the crowd is penalized by an additional loss.

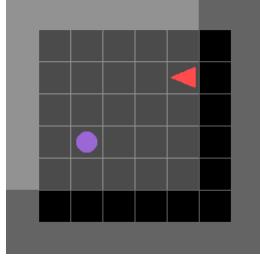
In Krueger (2016), Schulze and Evans (2018), active reinforcement learning (ARL) is proposed: different from usual RL, the agent has to choose in each round if it *wants* to receive the reward that results from its action, which results in a constant query-cost $c > 0$. Note that in this setting, what is queried is feedback, whereas in our setting, the model queries guidance *prior* to making a decision. Active Reward Learning (Daniel et al., 2014) is a similar approach in the context of continuous control tasks.

3 Approach

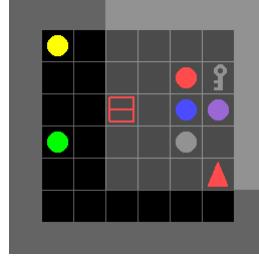
3.1 BabyAI Game

All our experiments take place in the BabyAI platform (Chevalier-Boisvert et al., 2019). In this platform, an agent learns to complete tasks given by instructions in a subset of English in a mini grid-world environment. The environments are only partially observable to the agent.

Figure 1 shows two example levels. In total there are 19 different levels that increase in difficulty and complexity of tasks. For each level, the BabyAI framework can randomly generate many missions that require roughly the same skillset and are provided with a similar language instruction. For the following investigation, we only focus on the levels “GoToObj” and “PutNextLocal”. These are chosen to be simple but nevertheless require a representative set of skills. As such, our results can be understood as a proof of concept.



(a) GoToObj



(b) PutNextLocal

Figure 1: Example levels from BabyAI. GoToObj is the simplest level, requiring the agent to go to a specific object. This task is given by a language instruction to the agent, such as “Go to the purple ball”. PutNextLocal is more difficult, requiring a more complex skill-set. An example task in this setting would be “Put the grey key next to the red box”. Shown in the images are the agent (red triangle), different objects (boxes, keys, balls in different colors) and the observation of the agent (a brighter 7×7 visual field that has the agent in the middle of the bottom row relative to the agent. Parts of this is outside of the shown images)

3.2 Model

In this section, we describe the model that we use for a learner that may ask for guidance. We refer to Figure 2 for a visual explanation.

Based on the observation o_t , the instruction i and its own memory unit, the learner L builds a representation r_t :

$$r_t = L(o_t, i). \quad (1)$$

o_t is a $7 \times 7 \times 3$ tensor describing the viewable environment (e.g. the brighter area in Figure 1) and i is a natural language instruction (e.g. “Go to the purple ball”). The representation unit L uses a FiLM module (Perez et al., 2017) followed by a memory RNN.

First, consider a learner without guidance. In this setting it directly takes the representation r_t and feeds it into the policy module P that outputs the next action $a_t = P(r_t)$. For more details on this baseline setting, see Chevalier-Boisvert et al. (2019).

A guided learner

Now consider Mul et al. (2019), where a guide is added. The guide follows the same architecture as the learner-policy combination, but between the representing unit and the policy there is a discrete “bottleneck” that only allows 9 different messages to pass. The policy then needs to encode this message continuously in order to choose the correct

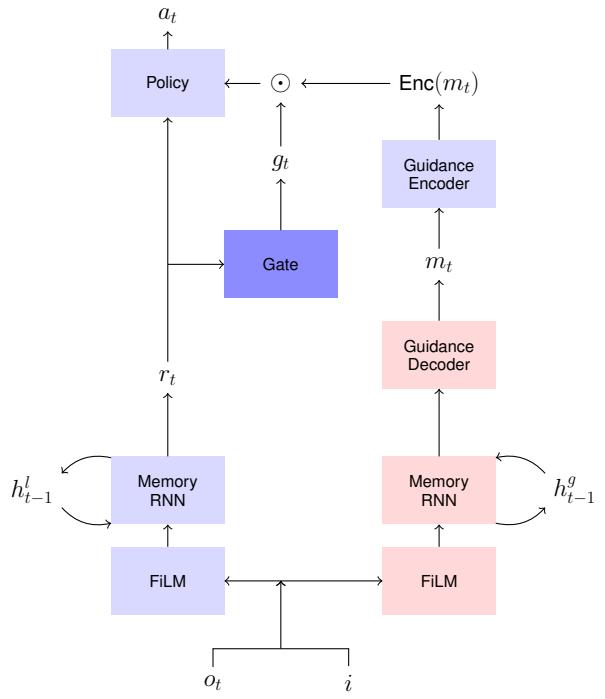


Figure 2: Architecture of a learner that can ask for guidance. Depicted variables are o_t : observation input, i : linguistic instruction, h_{t-1}^l and h_{t-1}^g : memory, r_t : learned representation, m_t : the discrete guidance message, g_t : the gating weight and a_t : the action chosen based on r_t and possibly the encoded message $\text{Enc}(m_t)$. The red part (the guide) is pretrained and then finetuned, while the blue parts (conceptually belonging to the learner) are newly initialized at the beginning of the training.

action out of 7 possibilities. After this guide-policy combination is trained, the messages are fed into the policy attached to a newly initialized learner in order to help it make its decision. In this later guided training stage, the policy of the guide is not used anymore.

More formally, the guide uses the same input and a memory unit to produce a message m_t of two words with 3 possible tokens for each.

$$m_t = G(o_t, i). \quad (2)$$

The message m_t is then encoded to a higher dimensional continuous encoding $\text{Enc}(m_t)$ that is produced by an encoder of the same architecture as the encoder used while training the guide. The policy then bases its decision on both the learned representation r_t and encoding $\text{Enc}(m_t)$, which are simply concatenated:

$$a_t = P(r_t, \text{Enc}(m_t)). \quad (3)$$

More details can be found in Mul et al. (2019).

Adding a gate

To enable the learner to decide when to receive guidance, we extend the learner with a gate module G to learn a gating weight $g_t \in \{0, 1\}$ that switches the policy input between $(r_t, \text{Enc}(m_t))$ (guided) and $(r_t, \mathbf{0})$ (unguided):

$$\begin{aligned} a_t &= P(r_t, g_t \cdot \text{Enc}(m_t)) \\ &= P((1 - g_t) \cdot (r_t, \mathbf{0}) + g_t \cdot (r_t, \text{Enc}(m_t))) \\ &= \begin{cases} P(r_t, \mathbf{0}), & \text{if } g_t = 0, \\ P(r_t, \text{Enc}(m_t)), & \text{if } g_t = 1, \end{cases} \end{aligned} \quad (4)$$

where $g_t = G(r_t)$. The gate module G is a two-layered MLP with a tanh activation functions that outputs a scalar, followed by a sigmoid activation function and a threshold function with parameter 0.5. The module G that produces the gating weight will from here on be referred to as the gate.

3.3 Training

We use a pretrained Reinforcement Learning (RL) expert, trained as in [Mul et al. \(2019\)](#) and [Chevalier-Boisvert et al. \(2019\)](#) by proximal policy optimization ([Schulman et al., 2017](#)). After training, the expert is placed once in many missions in order to create training data containing the expert behavior. Using imitation learning as in [Mul et al. \(2019\)](#), we have a cross-entropy loss function L_{ce} that measures how much the distribution over actions given by the policy of our model deviates from the “correct” action of the RL expert.¹ Furthermore, we penalize the learner for asking for guidance by adding the gating weight to the loss and balance these incentives by a hyperparameter λ :

$$L = L_{\text{ce}} + \lambda \cdot g. \quad (5)$$

We use $\lambda = 0.3$ in all GoToObj experiments and 0.05 for PutNextLocal, values that were found by hyperparameter search. The combined model consisting of pre-trained guide, also trained by imitation learning as in [Mul et al. \(2019\)](#), and the newly initialized learner is then trained end-to-end by backpropagating the gradients to all the weights

¹For mitigating confusion, we mention explicitly that the RL expert is not the same as the guide: the expert creates the data that is used for backpropagating the model and thus for training it *following* the choice of the action. The guide, however, gives its guidance *prior to* the decision about the chosen action.

of the combined model. In order to pass the gradients also through the discrete gate G , we use the straight-through estimator ([Bengio et al., 2013](#); [Courbariaux et al., 2016](#)). In order to allow the learner to learn the usefulness of the guidance at the beginning of training, we initialize G with a positive bias.

4 Experiments

In this section, we describe the experiments conducted in order to test the setting of a learner that queries the guide for help. In order to assess this, we train the combined model with $\lambda = 0.3$ for 7 runs on the simplest level, GoToObj, until convergence. In this level, the learner is instructed to go to a specific object. Results with $\lambda = 0.05$ for 7 runs on the level PutNextLocal can be found in Appendix B.

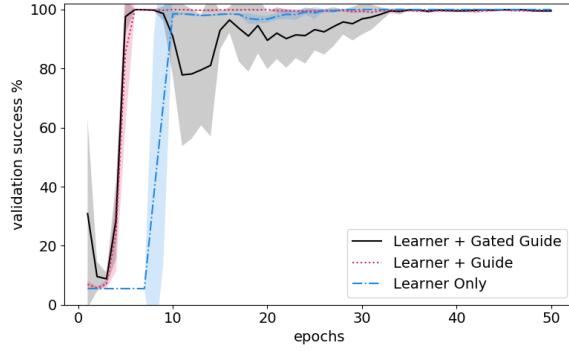
Performance and dynamics. First of all, we are interested in how our model performs compared to baselines. The first baseline is the learner on its own trained with imitation learning, which was the setting in [Chevalier-Boisvert et al. \(2019\)](#). The second baseline is the learner that receives guidance in each round without the need to query for it, which was studied in [Mul et al. \(2019\)](#). We expect that our model learns faster than the original learner model, due to its access to guidance, but slower than the guided learner without the gate, since it does not always receive guidance. The results can be found in Figure 3 (a), where we plot the success rate, i.e. portion of successfully completed episodes, during validation.

We monitor the average gating weight over epochs, which we call *guidance rate*, to inspect the usage of the gate over the course of the training. Furthermore, we compare the overall accuracy with the accuracy conditioned on the cases of an open or closed gate to assess the influence on performance. These metrics can be found in Figure 3 (b)².

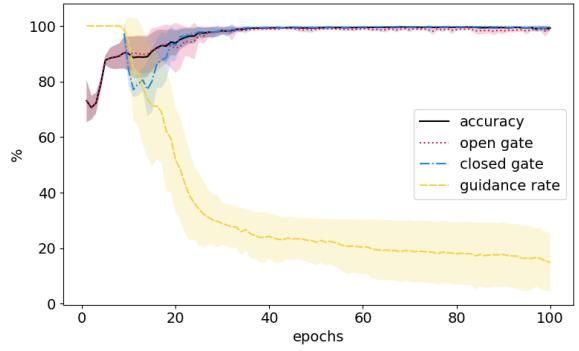
Since the accuracy plots in Figure 3 (b) are solely correlational, we furthermore plot the validation accuracy for the case where we *intervene* during validation in the gate in order to have it opened or closed to assess the causal influence of the open gate on the accuracy, see Figure 4.

Economic requests. While the accuracy mea-

²The accuracy is the percentage of chosen actions that coincide with the “correct” action of the RL expert that is used in the imitation learning process.



(a) Baseline Success Rate Comparison



(b) Validation Accuracy Comparison

Figure 3: Results of training our combined model until convergence on GoToObj. Results are averaged over 7 runs. Shaded regions represent standard deviations.

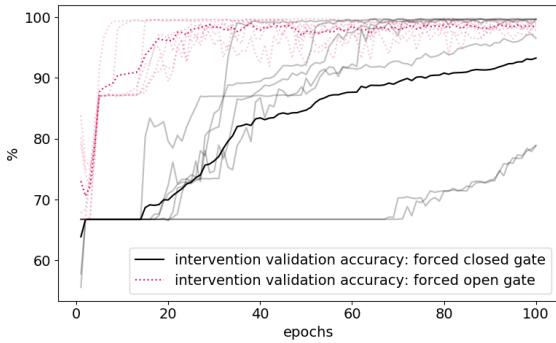


Figure 4: We compare the accuracy during validation in cases of forced open and closed gates: irrespective of the gating weight g_t computed from the system, we set $g_t = 1$ (so that the policy bases its decision on the encoded guidance $\text{Enc}(m_t)$) for the red dotted curve and $g_t = 0$ for the black curve.

sures how often the agent was “right”, the cross-entropy policy loss gives greater insights into the performance with respect to the actual training objective. We would like to assess whether the learner uses the gate economically, since it is penalized. This means to ask for guidance in situations in which it expects the greatest reduction in the policy loss. The policy loss for cases of open and closed gate can therefore be found in Figure 5. We compare it with the counterfactual policy losses that arise if we force the gate to be opened if the learner wants it to be closed and vice versa. This intervention now allows us to assess the *causal* influence of the gate on the policy loss.

Guidance semantics. Finally, we are interested in whether there are meaningful correlations between the frames in which the learner asks for guidance and the actions that the learner takes

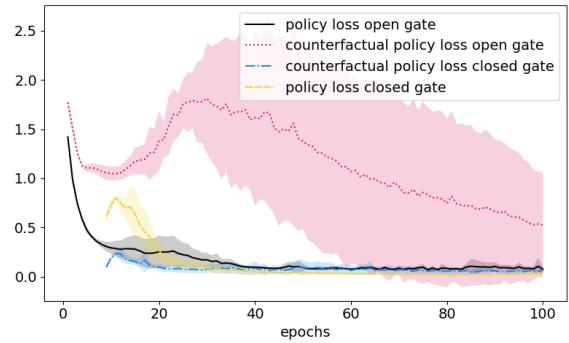


Figure 5: Policy Loss Comparison

(Figure 6) as well as the messages emitted by the guide (Figure 7). By analyzing this, we can see if the learner masters certain situations that require a certain action or are accompanied by a certain message.

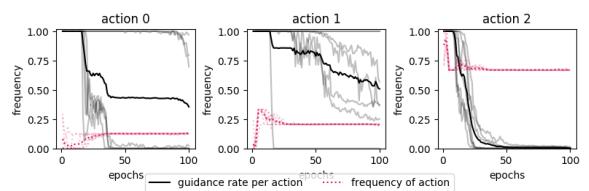


Figure 6: Frequency of open gate conditioned on actions and frequencies of actions themselves.

5 Analysis

5.1 Results and analysis of experiments

After outlining the experiments, we now briefly analyze the results.

Performance and dynamics. First, we notice in Figure 3 (a) that, while initially our model is

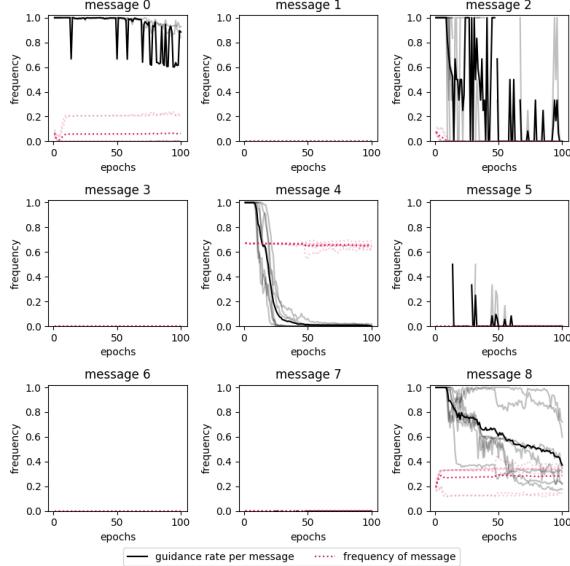


Figure 7: Frequency of open gate conditioned on messages and frequencies of messages themselves.

indeed faster than the learner baseline and keeps up with the guided baseline, from epoch 10 to 30 there is a performance dip not seen in the former models. This is precisely the phase in which the gate successively closes more and more, as can be seen in Figure 3 (b). Our explanation for this dip is that the policy first needs to adjust to the fact that it does not get the familiar guidance anymore. Eventually, from epoch 30 onward, our model performs almost perfectly and as well as the baselines.

As mentioned already, the learner becomes indeed more independent over time, roughly choosing its action on its own in 80% of the cases from epoch 50 onward. As soon as the guidance rate begins to drop in epoch 10, we can compute an accuracy conditioned on cases where the gate is closed, as can be seen in the blue line of Figure 3 (b). This accuracy is lower than the accuracy in case of an open gate (red dotted line) and only fully catches up roughly after 25 epochs of training.

About the intervention accuracy in Figure 4: we observe that the initial phase with a guidance rate at 1 sees a steep increase in accuracy with guidance but nearly no change in the accuracy without guidance. We suppose that is the case since in this phase the training happens exclusively with an open gate. Then, the guidance rate drops and training happens increasingly without guidance. Accordingly, the accuracy without guidance starts to increase and eventually catches up. In many individual runs, the performance without guidance

is ultimately actually even better. We hypothesize that this is since the gate is mostly closed and so the policy doesn't "expect" the gate to be open anymore. Consequently, an open gate and additional encoded message is confusing and leads to misbehavior. Intuitively, this is similar to how humans who are very experienced in their profession may actually just be distracted by someone who occasionally tries to give them advice instead of just letting them do their task on their own.

Economic requests. In order to get a better feeling for how smart the agent is in its guidance requests, we look at Figure 5. For similar results about the entropy, see Appendix A. We see an overall tendency for the policy loss to decrease, as we would expect due to the training. At the same time, the situations where the gate is open are those that are more difficult for the learner (including in the comparison of the counterfactual cases where we change the gate). Additionally, in those situations the reduction in policy loss achieved by asking for guidance is greater – this can be seen by comparing with the counterfactual situations. We furthermore observe that after the guidance rate starts to drop around epoch 10, the policy loss in situations of a closed gate rapidly sinks as the learner adapts to those novel situations. In the meantime, the open gate policy loss stabilizes until around epoch 20, while the counterfactual policy loss in these situations strongly increases. This indicates that the learner learns to selectively open the gate in situations that are more difficult and especially so without guidance.

Guidance semantics. To gain insights about dependencies between specific actions and the guidance rate, we now look at Figure 6. We see that in situations where the learner takes action 2, which corresponds to "move forward", the guidance rate drops relatively early to 0. This may be the case since this is the most common and supposedly most easily identifiable action. For action 0 that corresponds to "turn right" and action 1 that corresponds to "turn left", we see that the guidance rate also decreases, albeit slower and asymmetrically. This may be due to a higher difficulty of distinguishing those actions from each other. Intuitively they are symmetric and it may often be unclear what to do if "move forward" is not a promising action. In some runs, the guidance rate drops more for action 0 and in some more for action 1. We may attempt to explain this by the learner ei-

ther learning to request help in situations where action 0 is one of the promising options (potentially the only one) or learning the same with action 1. In both cases, it is ensured that situations with confusion between those two actions are encompassed.

On the guidance rate conditioned on messages: as we can see in Figure 7, mainly three messages are used to convey guidance and for all of them the guidance rate decreases over time. We suppose that the overall trends happen due to the close correspondence between messages and actions that was already observed by [Mul et al. \(2019\)](#).

5.2 Guidance in space and time

So far, we mainly discussed “global” metrics, in the sense that we aggregate information over complete epochs. This still leaves open the question how guidance requests evolve with respect to the position of the agent and temporally during an episode.

For the first aim, we create heatmaps as in Figure 8. For more maps, see Figure 13 in Appendix A.

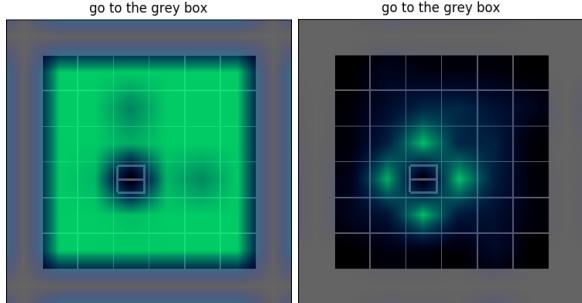


Figure 8: Two example heatmaps from training in level GoToObj, one in the beginning of training and one in an advanced stage. These heatmaps are created as follows: after an epoch is finished, the agent is placed in a specific mission. Then, we let it follow its path until the episode is over. For each point in the agent trajectory we record whether it asks for guidance. Multiple trajectories are sampled by randomly placing the agent in a new position. The brightness of the color in the figure depicts the average guidance rate within that position.

As we can see, the trained agent asks for guidance often specifically if it is near the goal object in order to find out if it should “turn towards it”, which would cause the goal to be reached. It is important for the agent to be reasonably sure about the goal being reached beforehand, since otherwise turning to the object will result in two lost moves.

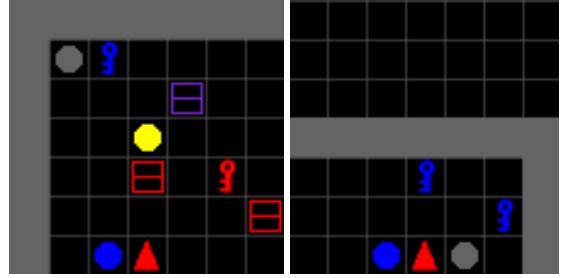


Figure 9: Observation types in GoToLocal. GoToLocal is a level different from GoToObj or PutNextLocal and is used to illustrate some of the extra possibilities $((1,x)$ and $(x,1)$) on further levels. In the left mission, the agent is tasked to go to the blue ball. This is directly left from the agent, whereas to the right, there is no feature in common with the goal. Therefore, the observation type is $(2, 0)$. In the right mission, the goal is to go to a grey ball. Since the blue ball shares one feature with the grey one (namely being a ball), the corresponding observation type is $(1, 2)$.

In order to test the influence of objects on the agent more quantitatively, we create another metric that conditions the guidance rate on how “goal-like” the observation is that the agent faces. For this matter, we assign tuples (d_1, d_2) to each observation, where d_1 and d_2 signify how goal-like the object left and right from the agent is. We measure this by the number of features it has in common with the goal object, where features are both color and object-type. $d_i = k$ means that the object shares $k \in \{0, 1, 2\}$ features with the goal object. This creates 9 observation types³. See Figure 9 for examples. The results can be found in Figure 10. We can observe strong changes in the guidance rate if the goal object is to the left or right: if it is on the left, the guidance rate is significantly greater and if it is to the right, the guidance rate is significantly smaller than usually. This is in line with the plots of the guidance rate conditioned on actions, Figure 6, which already showed that turning to the left requires considerably more guidance than turning to the right. This indicates that the high guidance rate at goal objects may to a large extend be caused by the high correlation between turning actions and guidance rates and be mostly independent of the fact that there is a goal

³Note that even the combination $(2, 2)$ is in principle possible in higher levels: There are tasks such as “Go to a red ball” where several red ball can be in the mission. However, this is unlikely and the expert never places itself between two goal objects. Furthermore, in GoToObj there is simply just one object in the mission. Therefore, the graph in Figure 10 corresponding to $(2, 2)$ is empty.

around.

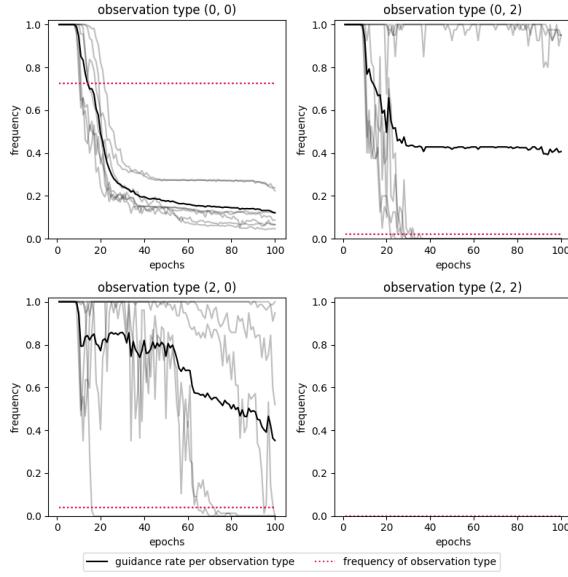


Figure 10: Frequency of open gate conditioned on observation types and frequencies of observation types themselves. For example, type $(2, 0)$ is a situation where directly left of the agent there is the goal and right of it there is no object sharing any feature with the goal-object. In GoToObj there is only one object in the level so situations like $(1, x)$ or $(x, 1)$ do not occur and are left out in this plot.

Now we turn to the question about the guidance rate in time: within one episode, are there usually phases where more or less guidance is needed? The heatmaps suggest that the agent mostly asks for guidance in the end of the episode.

In order to answer this question, we create the “guidance per time quantile” plot, Figure 11. As we can see, the guidance rate is in general high in the beginning of episodes and drops once more knowledge about the environment is acquired. However, in the end of the episode, the guidance rate grows again and is greatest in the very end, which is in line with the qualitative assessment from the heatmaps.

One interpretation for this is the following: in the beginning, the agent needs to roughly figure out “in which direction to head”. Once this is clear, it can walk there without further guidance. But in the very end, it needs more precision and asks for guidance again in order to finally find its goal. This is similar to how humans often look at a map in the beginning of a hike in order to figure out the direction, and then in the end again in order to reassess how their new position now relates to the goal.

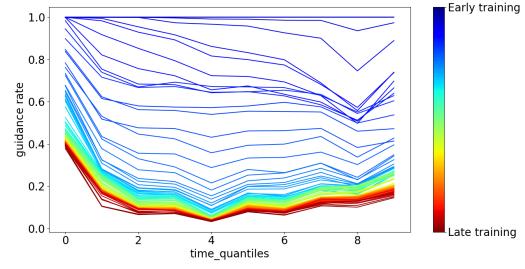


Figure 11: Guidance per time quantile: roughly speaking, a timepoint t is in quantile k of 10 if $t/l \approx k/10$, where l is the length of the corresponding episode. The plots show the guidance rate corresponding to the different quantiles. Dark blue curves belong to earlier epochs whereas red curves belong to later epochs.

6 Conclusion and future research

In this paper, we extended a recently proposed method to guide a learner in a gridworld environment by letting the learner explicitly ask for help. To accomplish this, we defined a binary gate in the learner’s model. We brought the original approach closer to the real world by (i) enabling bi-directional communication and (ii) attaching a cost to it.

We showed that the learner successfully learns to utilize the guidance gate to achieve a favorable trade-off between learning speed and amount of guidance requested. Initially using the full guidance to learn faster than a learner without guidance, it eventually learns to request guidance only where it is especially helpful, acting increasingly independent.

Future research could consist of retraining the guide to see if it learns to send more abstract messages that provide guidance over multiple time steps. This may be fruitfully combined with giving the guide an additional information advantage like a bird’s eye’s view so that the guide has more foresight than the learner. It would require a way for the learner to memorize past messages.

Another direction is to replace the gate by an emergent communication channel from the learner to the guide, so that the learner can send its guidance requests in more nuanced ways. Furthermore, we saw that the policy may have problems dealing with the additional guidance it receives unexpectedly in late training. It may be worthwhile to experiment with policy architectures that can deal better with spontaneous changes in its input.

Finally, research might as well aim at finding ways to meaningfully replace the guide agent by a human. This might allow for better learning in tasks that autonomous agents struggle to learn by themselves.

References

- Dario Amodei and Jack Clark. 2016. Faulty reward functions in the wild. <https://openai.com/blog/faulty-reward-functions/>. Accessed: 2019-06-22.
- Yoav Artzi and Luke Zettlemoyer. 2013. Weakly supervised learning of semantic parsers for mapping instructions to actions. *Transactions of the Association for Computational Linguistics*, 1:49–62.
- Yoshua Bengio, Nicholas Léonard, and Aaron Courville. 2013. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*.
- Maxime Chevalier-Boisvert, Dzmitry Bahdanau, Salem Lahiou, Lucas Willems, Chitwan Saharia, Thien Huu Nguyen, and Yoshua Bengio. 2019. BabyAI: First steps towards grounded language learning with a human in the loop. In *International Conference on Learning Representations*.
- J. Clouse. 1997. On integrating apprentice learning and reinforcement learning title2:. Technical report, Amherst, MA, USA.
- John D Co-Reyes, Abhishek Gupta, Suvansh Sanjeev, Nick Altieri, John DeNero, Pieter Abbeel, and Sergey Levine. 2018. Guiding policies with language via meta-learning. *arXiv preprint arXiv:1811.07882*.
- Matthieu Courbariaux, Itay Hubara, Daniel Soudry, Ran El-Yaniv, and Y Bengio. 2016. Binarized neural networks: Training deep neural networks with weights and activations constrained to +1 or -1.
- C. Daniel, M. Viering, J. Metz, O. Kroemer, and J. Peters. 2014. Active reward learning. In *Proceedings of Robotics: Science & Systems*.
- Serhii Havrylov and Ivan Titov. 2017. Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. In *Advances in neural information processing systems*, pages 2149–2159.
- Jiechuan Jiang and Zongqing Lu. 2018. Learning attentional communication for multi-agent cooperation. In *Advances in Neural Information Processing Systems*, pages 7254–7264.
- Joohyun Kim and Raymond J Mooney. 2012. Unsupervised pcfg induction for grounded language learning with highly ambiguous supervision. In *Proceedings of the 2012 Joint Conference on Empirical Methods* *in Natural Language Processing and Computational Natural Language Learning*, pages 433–444. Association for Computational Linguistics.
- Vanessa Kosoy. 2019. Delegative reinforcement learning: learning to avoid traps with a little help. *Safe Machine Learning workshop at ICLR*.
- David Krueger. 2016. Active reinforcement learning : Observing rewards at a cost.
- Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2016. Multi-agent cooperation and the emergence of (natural) language. *arXiv preprint arXiv:1612.07182*.
- Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. 2018. Scalable agent alignment via reward modeling: a research direction. *CoRR*, abs/1811.07871.
- Hongyuan Mei, Mohit Bansal, and Matthew R Walter. 2016. Listen, attend, and walk: Neural mapping of navigational instructions to action sequences. In *Thirtytieth AAAI Conference on Artificial Intelligence*.
- Tomas Mikolov, Armand Joulin, and Marco Baroni. 2016. A roadmap towards machine intelligence.
- Dipendra Misra, John Langford, and Yoav Artzi. 2017. Mapping instructions and visual observations to actions with reinforcement learning. *arXiv preprint arXiv:1704.08795*.
- Igor Mordatch and Pieter Abbeel. 2018. Emergence of grounded compositional language in multi-agent populations. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- Mathijs Mul, Diane Bouchacourt, and Elia Bruni. 2019. Mastering emergent language: learning to guide in simulated navigation. *arXiv e-prints*, page arXiv:1908.05135.
- Khanh Nguyen, Debadeepa Dey, Chris Brockett, and Bill Dolan. 2019. Vision-based navigation with language-based assistance via imitation learning with indirect intervention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12527–12537.
- Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. 2017. Film: Visual reasoning with a general conditioning layer.
- Dean A. Pomerleau. 1991. Efficient training of artificial neural networks for autonomous navigation. *Neural Computation*, 3:97.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *ArXiv*, abs/1707.06347.
- Sebastian Schulze and Owain Evans. 2018. Active reinforcement learning with monte-carlo tree search. *CoRR*, abs/1803.04926.

Keenon Werling, Arun Tejasvi Chaganty, Percy S Liang, and Christopher D Manning. 2015. [On-the-job learning with bayesian decision theory](#). In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3465–3473. Curran Associates, Inc.

Terry Winograd. 1971. Procedures as a representation for data in a computer program for understanding natural language. Technical report, MASSACHUSETTS INST OF TECH CAMBRIDGE PROJECT MAC.

Haonan Yu, Haichao Zhang, and Wei Xu. 2018. Interactive grounded language acquisition and generalization in a 2d world. *arXiv preprint arXiv:1802.01433*.

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.

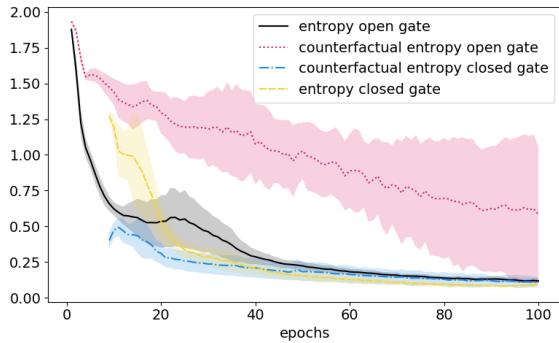


Figure 12: Entropy Comparison for GoToObj

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.

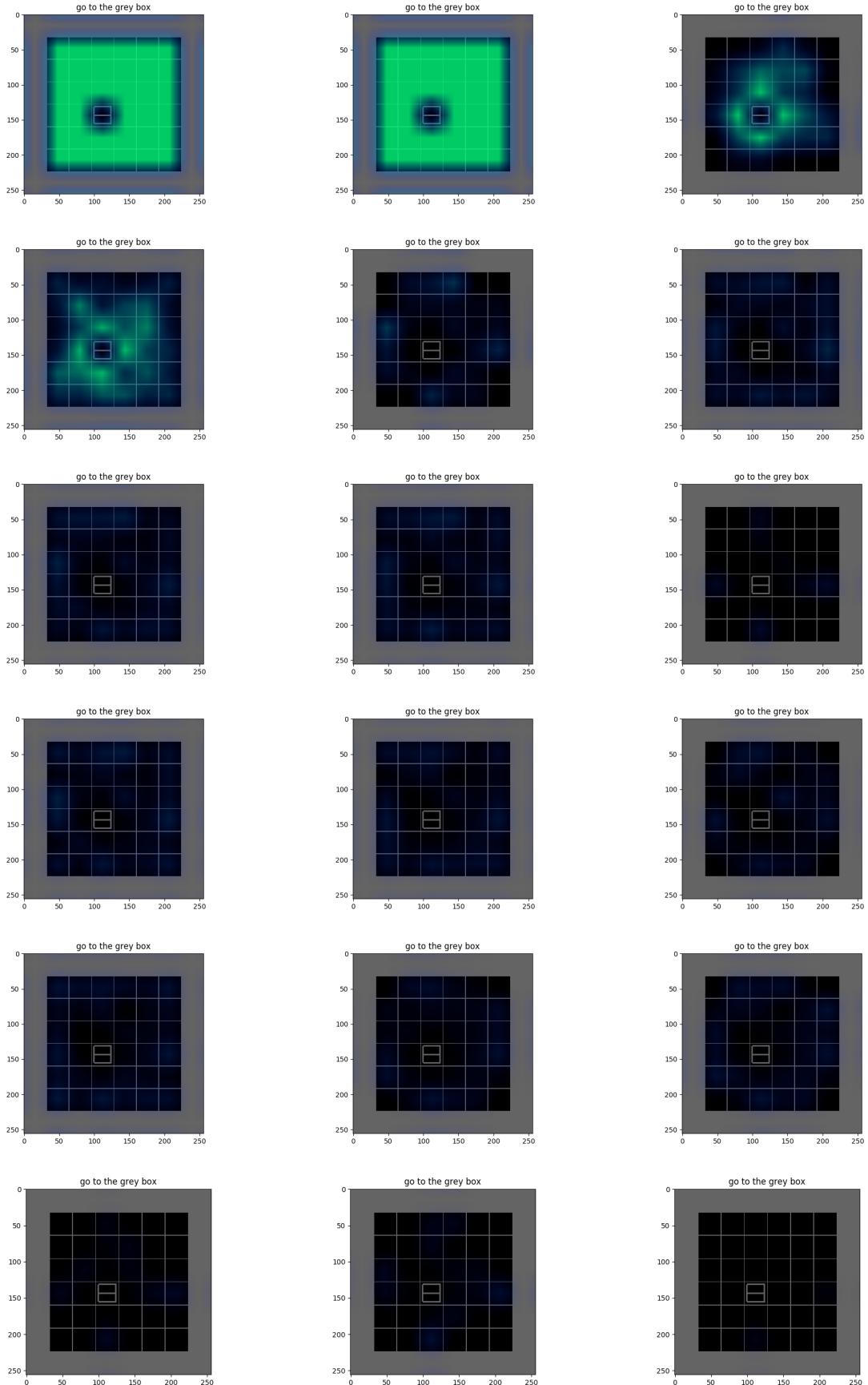


Figure 13: Heatmaps for GoToObj. They are ordered from left to right and then top to bottom. This shows how the guidance requests evolve over the course of the whole training in one specific example mission. Over time, not much guidance remains.

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.

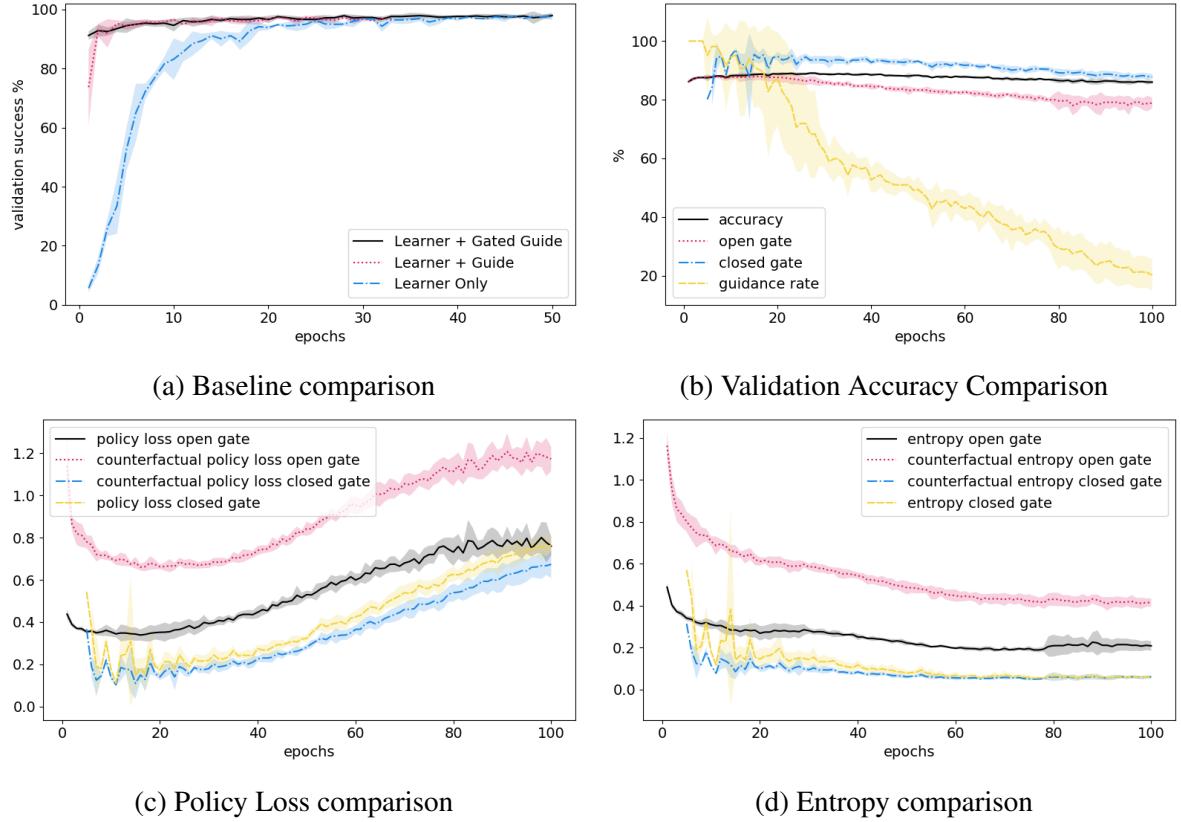


Figure 14: PutNextLocal

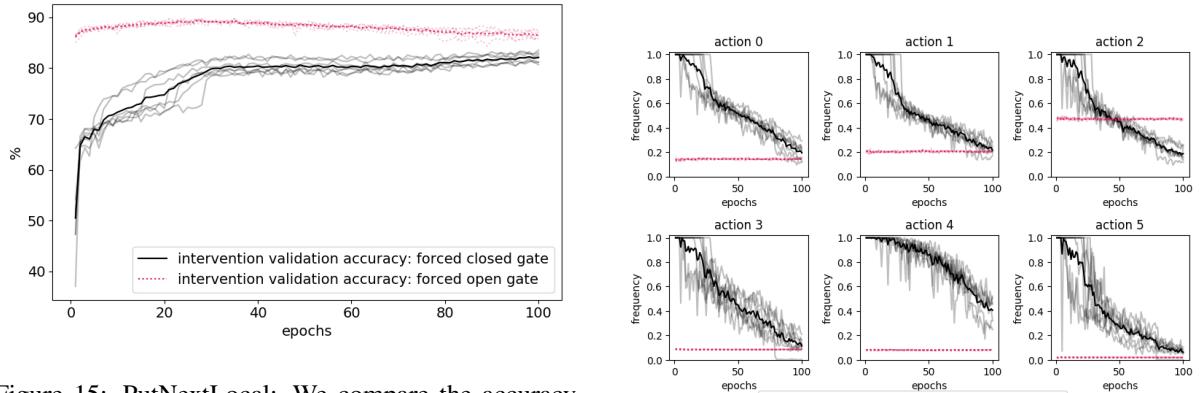


Figure 15: PutNextLocal: We compare the accuracy during validation in cases of forced open and closed gates: irrespective of the gating weight g_t computed from the system, we set $g_t = 1$ (so that the policy bases its decision on the encoded guidance $\text{Enc}(m_t)$) for the red dotted curve and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

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

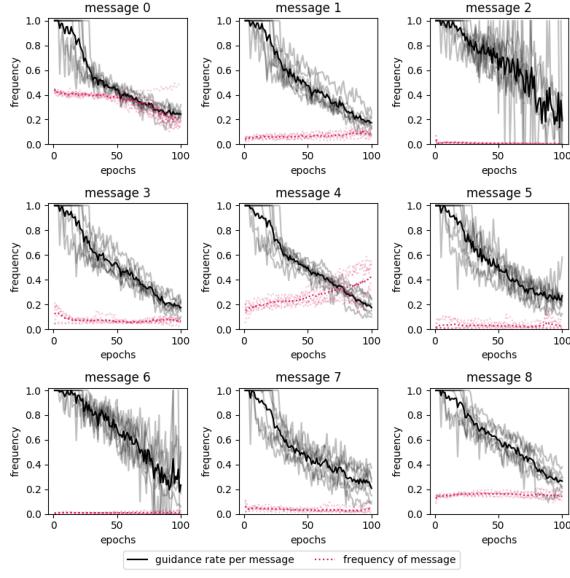


Figure 17: PutNextLocal: Frequency of open gate conditioned on messages and frequencies of messages themselves.

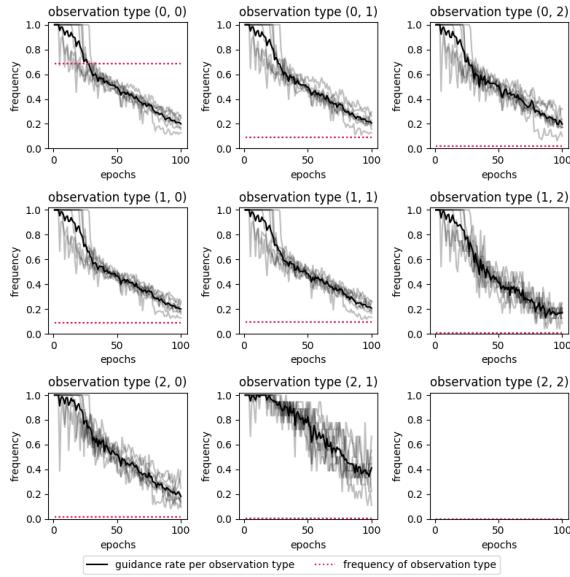


Figure 18: PutNextLocal: Frequency of open gate conditioned on observation types and frequencies of observation types themselves. For example, type (2, 1) is a situation where directly left of the agent there is the goal and right of it there is an object sharing one feature with the goal-object.

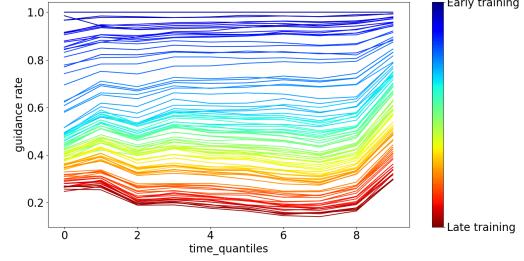


Figure 19: PutNextLocal: Guidance per time quantile: roughly speaking, a timepoint t is in quantile k of 10 if $t/l \approx k/10$, where l is the length of the corresponding episode. The plots show the guidance rate corresponding to the different quantiles. Dark blue curves belong to earlier epochs whereas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

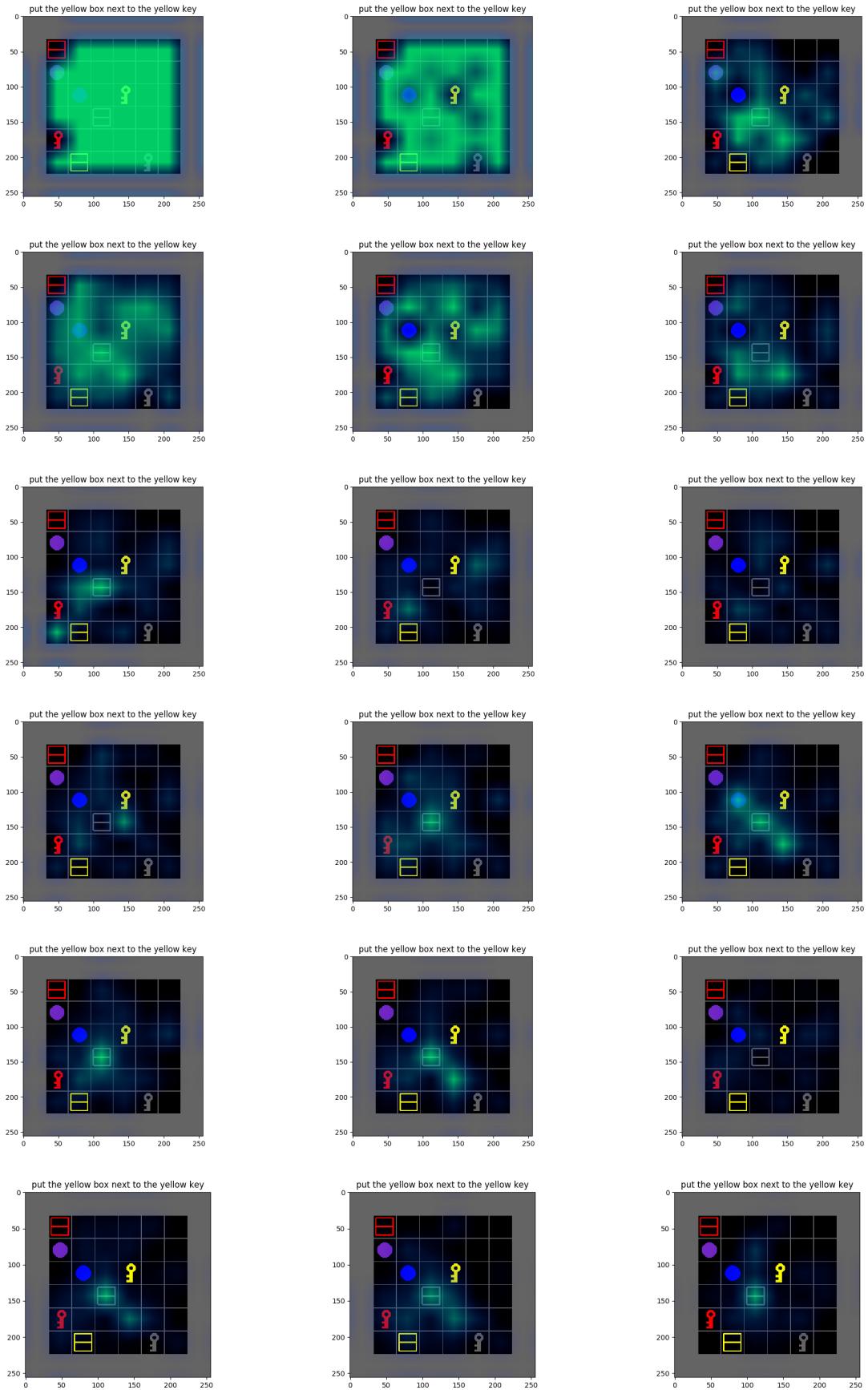


Figure 20: Heatmaps for PutNextLocal. They are ordered from left to right and then top to bottom. This shows how the guidance requests evolve over the course of the whole training in one specific example mission.