A Ablation with Fewer Objects

Table A1: **Performance of models on the symbolic version of 🦣IVRE with fewer objects and blickets.** We conduct experiments where at most 3 blickets are selected from 7 unique objects in 🦣IVRE. Results show that reducing the number of objects in 🦣IVRE would simplify the environment.

Model	Random	Bayes	Naive	Search-Naive	Search-Random	DDPG-FF	TD-3-FF
Acc	7.61	71.92	94.56	100.00	54.53	90.08	88.64
R	-12.81	4.73	10.82	14.32	3.17	11.96	11.53

Striking a balance between simplicity and complexity is crucial to maintaining the 🦣IVRE's ability to assess agents' reasoning abilities effectively. We conduct experiments where at most 3 blickets are selected from 7 different objects in 🦣IVRE. The results are shown in Tab. A1. As demonstrated, reducing the number of objects in 🦣IVRE would simplify the environment. However, this approach could inadvertently lead to shortcuts, as random or naive trials might efficiently address much of the uncertainty.

B Agent Details

B.1 RL Agent Details

All RL models are implemented in PyTorch (Paszke et al., 2017) under the helper library of Tianshou (Weng et al., 2022). When training the MLP and the LSTM backbones, we use the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 3×10^{-4}. The DDPG and the TD-3 agent share the same hyper-parameters: we set the exploration noise to 0.1 and the target network's soft update to 0.005. Other parameters have been set using the default parameters from the original work. All RL models were trained for 10^7 steps during training. The best model from training was saved and used during the evaluation phase. All the experiments reported herein were run on NVIDIA GeForce RTX 3090 or Tesla V100 graphics cards.

C Model Details

C.1 Symbol-Input Backbone

Tab. A2 shows the feed-forward architecture we use for the RL agents in the symbol-input version of the environment, where the parameter of the Linear layer denotes the size of its output and m is set equal to the size of the action space. Of note, the DDPG agent uses an actor network and a critic network, while the TD-3 agent uses one actor network and two critic networks. The actor network and the critic network share the same architecture. Note that the feed-forward architecture takes in both the current observation and the belief from the previous time step. Our online learning baseline PPO shares the same three-layer MLP for its backbone.

The recurrent architecture of the RL agent recruits the network components listed in Tab. A3, where we explicitly use a single-layer LSTM with a hidden size of 512 units.

Table A2: Feed-forward architecture of RL agents.

Layer	Units
Linear	512
ReLU	/
Linear	512
ReLU	/
Linear	512
ReLU	/
Linear	m

Table A3: Recurrent architecture of RL agents.

Layer	Units
LSTM	512
Linear	512
ReLU	/
Linear	m

C.2 Pixel-Input Backbone

The pixel-input model leverages a pre-trained ResNet-18 (He et al., 2016) for extracting image features. Specifically, we resize each image rendered from the Blender `EEVEE` engine (Blender Online Community, 2016) into 224 × 224 and feed it through the backbone. The features are then passed through the same actor network and the critic network listed in Tabs. A2 and A3.

D LLMs Experiments

We use the OpenAI Chat Completions API for our LLMs experiments; see examples in Tab. A4.

Table A4: **An example dialogue generated between GPT and our `IVRE` (the *User*).**

Role	Content
User	"There are nine objects: 1, 2, 3, 4, 5, 6, 7, 8, 9 that can be put on a machine. Up to four of all the objects have a unique property that can activate the machine. Your goal is to find out all the objects that can activate the machine in ten rounds. All the following questions in the format of '1, 2, and 3.'"
User	"Round 0: 1 are on the machine, and the machine is not activated. Question: Which objects can activate the machine?" "Answer: From 1-9, objects that can activate the machine are"
Assistant	"2, 3, 4, 5, 6, "
User	"Your answer is wrong."
User	"Round 1: 3 are on the machine, and the machine is not activated. Question: Which objects can activate the machine? Answer: From 1-9, objects that can activate the machine are"
Assistant	"2, 4, 5, 6, 7, "
User	"Your answer is wrong."
User	"Round 2: 1, 6, 8 are on the machine, and the machine is not activated. Question: Which objects can activate the machine? Answer: From 1-9, objects that can activate the machine are "
Assistant	"2, 3, 4, 5, 7, "
User	"Your answer is wrong."
User	"Round 3: 3, 4, 7, 9 are on the machine, and the machine is activated. Question: Which objects can activate the machine? Answer: From 1-9, objects that can activate the machine are "
Assistant	"3, 4, 7, and 9."
User	"Your answer is wrong."
User	"Round 4: None are on the machine, and the machine is not activated. Question: Which objects can activate the machine? Answer: From 1-9, objects that can activate the machine are "
Assistant	"Please provide the objects on the machine in Round 4 for me to give..."
User	"Now design your own trial, you can put some of the objects on the machine to validate your hypothesis. Question: Which objects will you put on the machine? Answer: From 1-9, I will put "
Assistant	"2, 5, and 6 on the machine."
User	"Your answer is wrong."
User	"Round 5: 2, 5, 6 are on the machine, and the machine is not activated. Question: Which objects can activate the machine? Answer: From 1-9, objects that can activate the machine are "

Assistant	"3, 4, 7, and 9."
User	"Now design your own trial, you can put some of the objects on the machine to validate your hypothesis. Question: Which objects will you put on the machine? Answer: From 1-9, I will put "
Assistant	"3 and 4 on the machine."
User	"Your answer is wrong."
User	"Round 6: 3, 4 are on the machine, and the machine is not activated. Question: Which objects can activate the machine? Answer: From 1-9, objects that can activate the machine are "
Assistant	"7 and 9."
User	"Now design your own trial, you can put some of the objects on the machine to validate your hypothesis. Question: Which objects will you put on the machine? Answer: From 1-9, I will put "
Assistant	"7 on the machine."
User	"Your answer is wrong."
User	"Round 7: 7 are on the machine, and the machine is not activated. Question: Which objects can activate the machine? Answer: From 1-9, objects that can activate the machine are "
Assistant	"9."

E Case Study

We give examples of interaction replay from different agents in 🦣`IVRE` in Figs. A1 to A7. In each case, the first two figures show available objects in an episode, where the first one contains Blickets and the second one non-Blickets. For the naive agent, it always proposes a trial composed of a single object (see Fig. A1), and from its result of activation, the agent deterministically assigns Blicketness to an object. A search-based naive agent proposes trials based on the confidence of every object. Figs. A3 and A4 show trials from such an agent. In Fig. A3, the agent is quite sure of the non-Blicketness of the purple and green metal cube after the context panels, so it tests other objects. However, it cannot figure out the Blicketness of the green metal ball through all of the contexts, leading to its failure.

The DDPG agent, however, has unique behavior. In Fig. A6, all the Blicket machines in initial contexts are activated, so the agent has no chance to generate correct belief from the context only, except guessing. However, the DDPG agent makes repetitive trials, making it harder to gain effective information to solve the problem. In follow-up experiments, the DDPG agent also solves tasks when most objects are non-Blickets. But it often fails when there are more Blickets, which means it cannot handle situations with high correlation.

Humans show powerful abilities in both reasoning and exploration. Fig. A7 shows an example where a successful trial is achieved. We can see that a flexible trial policy is adopted, and humans can reason effectively given limited contexts.

F Data Documentation

We follow the datasheet proposed in Gebru et al. (2021) for documenting our proposed benchmark:

1. **Motivation**
   (a) **For what purpose was the benchmark created?**
   The benchmark was created as an environment for evaluating artificial agents' reasoning ability under uncertainty. 🦣`IVRE` is an interactive environment featuring rich scenarios centered around Blicket detection. Agents in 🦣`IVRE` are placed into environments with various ambiguous action-effect pairs and asked to figure out each object's role. Agents are encouraged to propose effective and efficient experiments to validate their

hypotheses based on observations and gather more information. The game ends when all uncertainties are resolved or the maximum number of trials is consumed.

(b) **Who created the dataset and on behalf of which entity?**
This dataset was created by Manjie Xu, Guangyuan Jiang, Wei Liang, Chi Zhang and Yixin Zhu. They are from Beijing Institute of Technology (Manjie Xu, Wei Liang), Peking University (Guangyuan Jiang, Yixin Zhu) and National Key Laboratory of General Artificial Intelligence, BIGAI (Chi Zhang).

(c) **Who funded the creation of the dataset?**
M.X., G.J., W.L., C.Z., and Y.Z. are supported in part by the National Key R&D Program of China (2022ZD0114900), M.X. and W.L. are supported in part by the NSFC (62172043), and Y.Z. is in part by the Beijing Nova Program.

(d) **Any other Comments?**
None.

2. **Composition**

(a) **What do the instances that comprise the benchmark represent?**
The benchmark contains episodes in which agents are tasked to figure out which objects are Blickets.

(b) **How many instances are there in total?**
N/A. Each episode in 🦣`IVRE` is randomly sampled and can have infinite tasks.

(c) **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?**
🦣`IVRE` contains all possible instances.

(d) **What data does each instance consist of?**
In each episode, an agent is presented with several initial observations of various object combinations (referred to as *context*). The context alone is insufficient to solve Blicketness for *all* objects. Hence, in each following step (referred to as *trials*), the agent proposes a new experiment of a specific object combination and updates its belief of Blicketness based on the outcome of experiments.

(e) **Is there a label or target associated with each instance?**
Yes.

(f) **Is any information missing from individual instances?**
No.

(g) **Are relationships between individual instances made explicit?**
Yes.

(h) **Are there recommended data splits?**
No.

(i) **Are there any errors, sources of noise, or redundancies in the benchmark?**
No.

(j) **Is the benchmark self-contained, or does it link to or otherwise rely on external resources (*e.g.*, websites, tweets, other datasets)?**
Self-contained.

(k) **Does the benchmark contain data that might be considered confidential (*e.g.*, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)?**
No.

(l) **Does the benchmark contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?**
No.

(m) **Does the benchmark relate to people?**
No.

(n) **Does the benchmark identify any subpopulations (*e.g.*, by age, gender)?**
No.

(o) **Is it possible to identify individuals (*i.e.*, one or more natural persons), either directly or indirectly (*i.e.*, in combination with other data) from the dataset?**
No.

   (p) **Does the dataset contain data that might be considered sensitive in any way (*e.g.*, data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?**
   No.

   (q) **Any other comments?**
   None.

3. **Collection Process**

   (a) **How was the data associated with each instance acquired?**
   We render them using Blender.

   (b) **What mechanisms or procedures were used to collect the data (*e.g.*, hardware apparatus or sensor, manual human curation, software program, software API)?**
   In each episode, 🦣`IVRE` will randomly sample objects and blickets from a pool.

   (c) **If the dataset is a sample from a larger set, what was the sampling strategy (*e.g.*, deterministic, probabilistic with specific sampling probabilities)?**
   N/A.

   (d) **Who was involved in the data collection process (*e.g.*, students, crowdworkers, contractors) and how were they compensated (*e.g.*, how much were crowdworkers paid)?**
   Manjie Xu and Guangyuan Jiang wrote the generation code.

   (e) **Over what timeframe was the data collected?**
   N/A.

   (f) **Were any ethical review processes conducted (*e.g.*, by an institutional review board)?**
   The dataset raises no ethical concerns.

   (g) **Does the dataset relate to people?**
   No.

   (h) **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (*e.g.*, websites)?**
   N/A.

   (i) **Were the individuals in question notified about the data collection?**
   N/A.

   (j) **Did the individuals in question consent to the collection and use of their data?**
   N/A.

   (k) **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?**
   N/A.

   (l) **Has an analysis of the potential impact of the dataset and its use on data subjects (*e.g.*, a data protection impact analysis) been conducted?**
   Yes.

   (m) **Any other comments?**
   None.

4. **Preprocessing, Cleaning and Labeling**

   (a) **Was any preprocessing/cleaning/labeling of the data done (*e.g.*, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?**
   N/A.

   (b) **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (*e.g.*, to support unanticipated future uses)?**
   N/A.

   (c) **Is the software used to preprocess/clean/label the instances available?**
   N/A.

   (d) **Any other comments?**
   None.

5. **Uses**
    (a) **Has the dataset been used for any tasks already?**
    No, the dataset is newly proposed by us.
    (b) **Is there a repository that links to any or all papers or systems that use the dataset?**
    Yes, we provide the link to all related information on our project website.
    (c) **What (other) tasks could the dataset be used for?**
    This dataset could be used for other reserach topics like causal discovery, causal reasoning and active learning.
    (d) **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?**
    N/A.
    (e) **Are there tasks for which the dataset should not be used?**
    N/A.
    (f) **Any other comments?**
    None.
6. **Distribution**
    (a) **Will the dataset be distributed to third parties outside of the entity (*e.g.*, company, institution, organization) on behalf of which the dataset was created?**
    No.
    (b) **How will the dataset be distributed (*e.g.*, tarball on website, API, GitHub)?**
    🦣`IVRE` could be accessed on our project website.
    (c) **When will the dataset be distributed?**
    🦣`IVRE` has already been released.
    (d) **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?**
    We release our benchmark under CC BY-NC [1] license.
    (e) **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?**
    No.
    (f) **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?**
    No.
    (g) **Any other comments?**
    None.
7. **Maintenance**
    (a) **Who is supporting/hosting/maintaining the dataset?**
    Manjie Xu and Guangyuan Jiang are maintaining.
    (b) **How can the owner/curator/manager of the dataset be contacted (*e.g.*, email address)?**
    `manjietsu@bit.edu.cn`, `jgy@stu.pku.edu.cn`
    (c) **Is there an erratum?**
    Future erratum will be released through the website.
    (d) **Will the dataset be updated (*e.g.*, to correct labeling errors, add new instances, delete instances')?**
    Yes.
    (e) **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (*e.g.*, were individuals in question told that their data would be retained for a fixed period of time and then deleted)?**
    N/A. The dataset does not relate to people.
    (f) **Will older versions of the dataset continue to be supported/hosted/maintained?**
    Yes.

---

[1] `https://creativecommons.org/licenses/by-nc/4.0/`

(g) **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?**
Yes. We have released the source code as well as a licence on our project website. Future developments are welcome.

(h) **Any other comments?**
None.

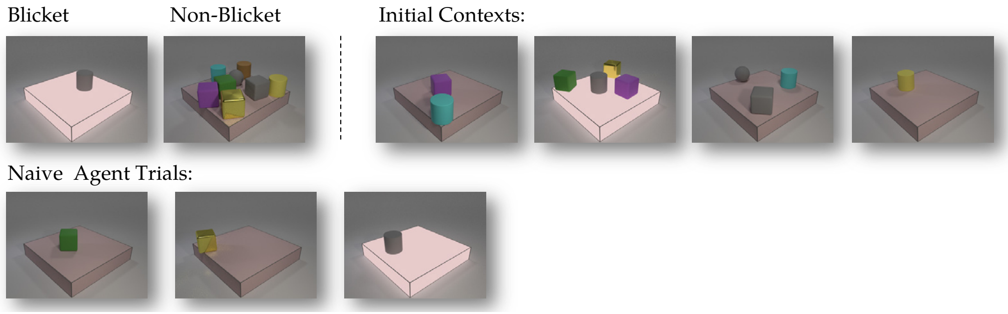


Figure A1: An example that is solved by the naive agent.

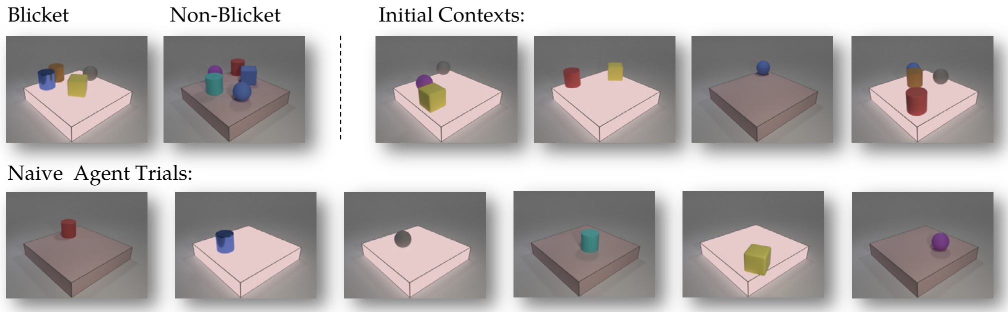


Figure A2: An example where the naive agent fails.

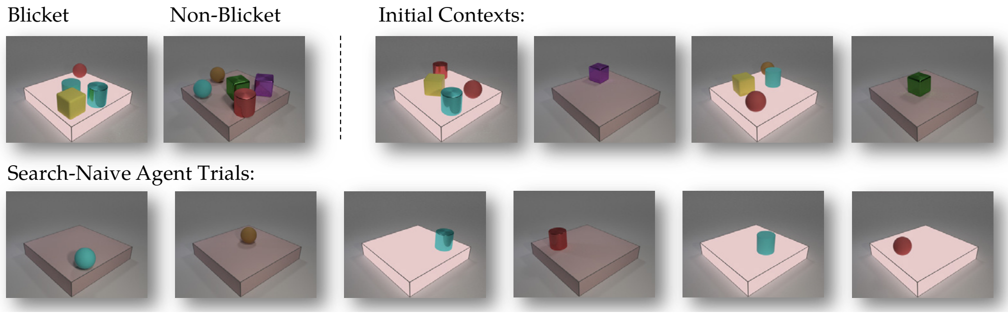


Figure A3: An example that is solved by the search-naive agent.

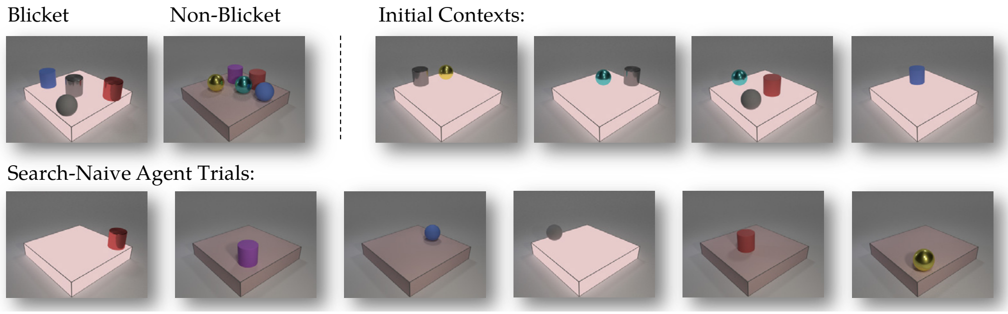


Figure A4: An example where the search-naive agent fails.

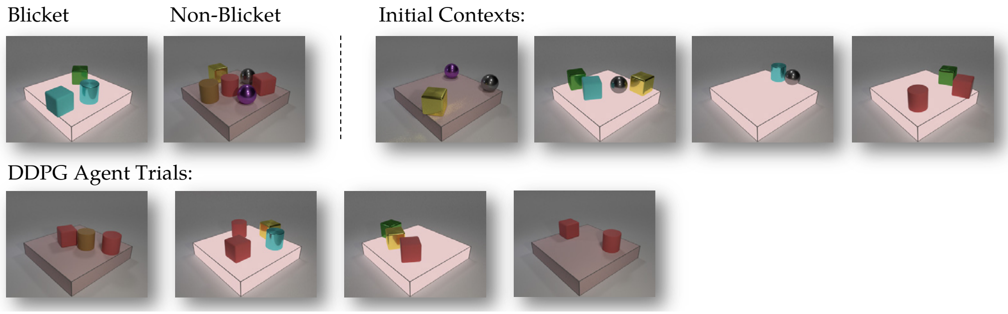


Figure A5: An example that is solved by the DDPG agent.

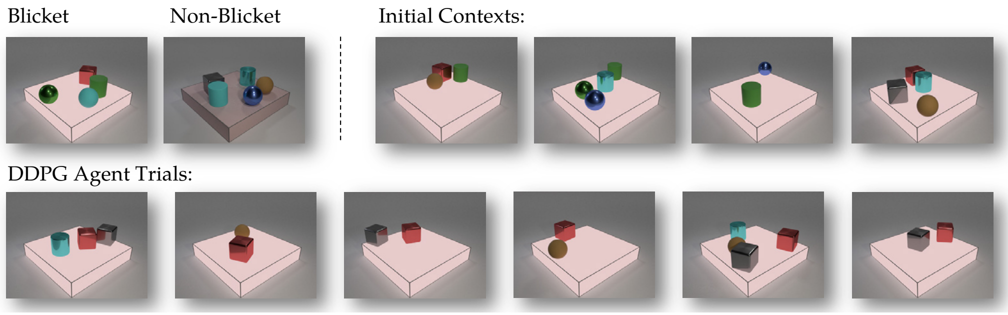


Figure A6: An example where the DDPG agent fails.

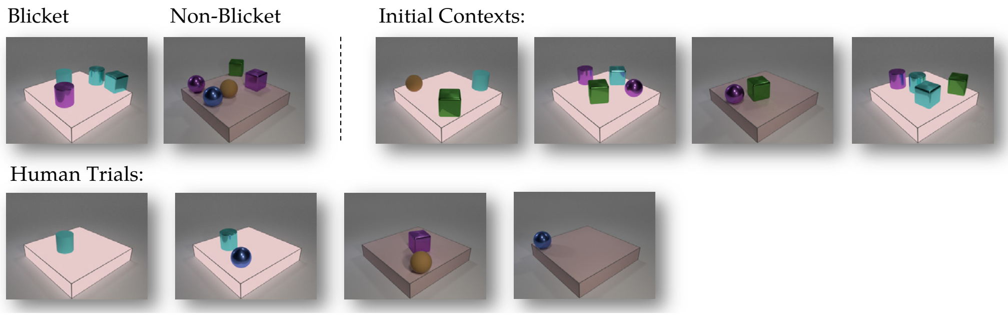
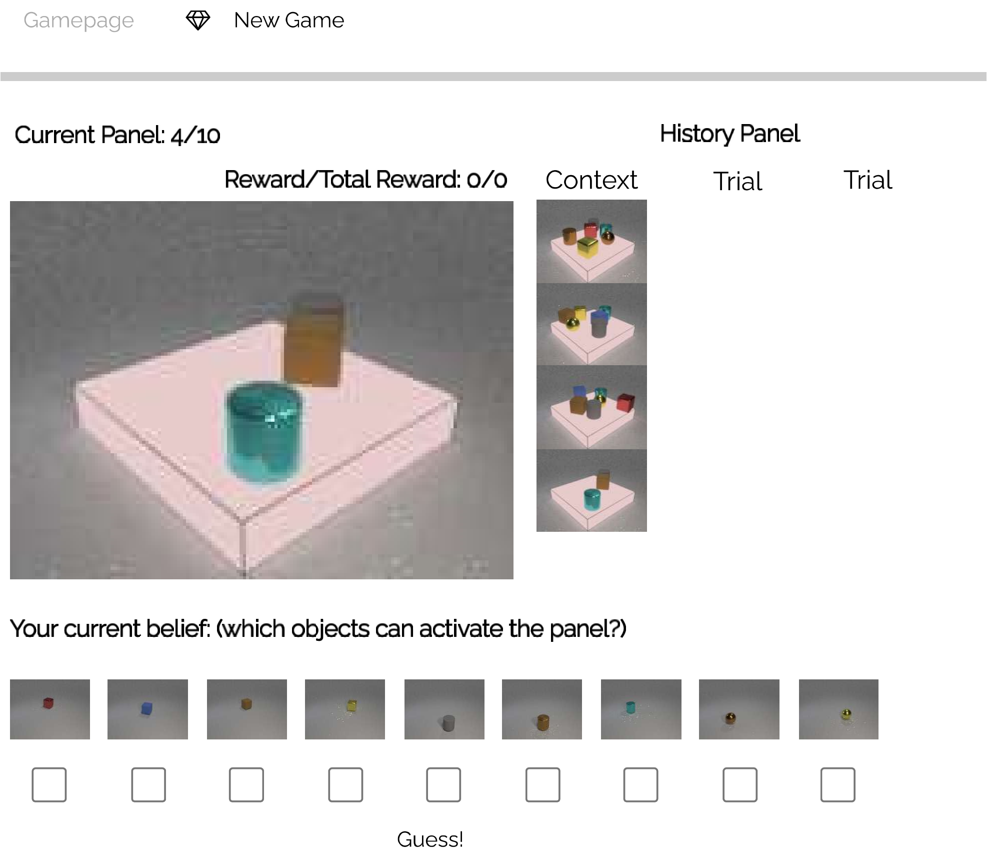


Figure A7: An example that is solved by a human participant.



Current Panel: 4/10

Reward/Total Reward: 0/0

Your current belief: (which objects can activate the panel?)

☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐

Figure A8: UI design for the web-based `IVRE` environment.