CRAFT: A Benchmark for Causal Reasoning About Forces and inTeractions

– Supplementary Material –

Tayfun Ates[1,*]
tates@hacettepe.edu.tr

M. Samil Atesoglu[1,*]
matesoglu@hacettepe.edu.tr

Cagatay Yigit[1,*]
cyigit@hacettepe.edu.tr

Ilker Kesen[2]
ikesen16@ku.edu.tr

Mert Kobas[3]
mkobas18@ku.edu.tr

Erkut Erdem[1]
erkut@hacettepe.edu.tr

Aykut Erdem[2]
aerdem@ku.edu.tr

Tilbe Goksun[3]
tgoksun@ku.edu.tr

Deniz Yuret[2]
dyuret@ku.edu.tr

[1] Hacettepe University Computer Vision Lab [2] Koç University Is Bank AI Center
[3] Koç University Language and Cognition Lab
https://sites.google.com/view/craft-benchmark

In this supplementary document, we provide additional details regarding our proposoed CRAFT dataset and our experimental evaluation. In particular, in Section A.1, we provide technical details about the functional modules we used to represent CRAFT questions. In Section A.2, we show sample functional programs for a number of questions to demonstrate how they are utilized in generating question and answer pairs. In Section A.3, we present results, showing predictions of top-performing three baselines on some sample questions of different type. In Section A.4, we give detailed descriptions about the human evaluation conducted to compare and contrast the performance differences between machine reasoning models and humans on CRAFT. Lastly, in Section B, we present the datasheet for our CRAFT dataset.

Contents

---

[*] indicates equal contributions.

Submitted to the 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks. Do not distribute.

A Appendix

A.1 Functional Modules

CRAFT questions are represented with functional programs. Input and output types for our functional modules are listed in Table A.1. Lists of all functional modules are also provided in Tables A.2-A.5.

Table A.1: Input and output types of functional modules in CRAFT.

Type	Description
Object	A dictionary holding static and dynamic attributes of an object
ObjectSet	A list of unique objects
ObjectSetList	A list of *ObjectSet*
Event	A dictionary holding information of a specific event
EventSet	A list of unique events
EventSetList	A list of *EventSet*
Size	A tag indicating the size of an object
Color	A tag indicating the color of an object
Shape	A tag indicating the shape of an object
Integer	Standard integer type
Bool	Standard boolean type
BoolList	A list of *Bool*

Table A.2: Input functional modules in CRAFT.

Module	Description	Input Types	Output Type
`SceneAtStart`	Returns the attributes of all objects at the start of the simulation	*None*	*ObjectSet*
`SceneAtEnd`	Returns the atttributes of all objects at the end of the simulation	*None*	*ObjectSet*
`StartSceneStep`	Returns `0`	*None*	*Integer*
`EndSceneStep`	Returns `-1`	*None*	*Integer*
`Events`	Returns all of the events happening between the start and the end of the simulation	*None*	*EventSet*

Table A.3: Output functional modules in CRAFT.

Module	Description	Input Types	Output Type
`QueryColor`	Returns the color of the input object	*Object*	*Color*
`QueryShape`	Returns the shape of the input object	*Object*	*Shape*
`Count`	Returns the size of the input list	*ObjectSet*	*Integer*
`Exist`	Returns true if the input list is not empty	*ObjectSet / EventSet*	*Bool*
`AnyFalse`	Returns true if there is at least one false in a boolean list	*BoolList*	*Bool*
`AnyTrue`	Returns true if there is at least one true in a boolean list	*BoolList*	*Bool*
`IsBefore`	Returns whether the first event happened before the second event	*(Event, Event)*	*Bool*
`IsAfter`	Returns whether the first event happened after the second event	*(Event, Event)*	*Bool*

Table A.4: Object filter functional modules in CRAFT.

Module	**Description**	**Input Types**	**Output Type**
`FilterColor`	Returns the list of objects which have a color same with the input color	*(ObjectSet, Color)*	*ObjectSet*
`FilterShape`	Returns the list ofobjects which have a shape same with the input shape	*(ObjectSet, Shape)*	*ObjectSet*
`FilterSize`	Returns the list of objects which have a size same with the input size	*(ObjectSet, Size)*	*ObjectSet*
`FilterDynamic`	Returns the list of dynamic objects from an object set	*ObjectSet*	*ObjectSet*
`FilterMoving`	Returns the list of objects that are in motion at the step specified	*(ObjectSet, Integer)*	*ObjectSet*
`FilterStationary`	Returns the list of objects that are stationary at the step specified	*(ObjectSet, Integer)*	*ObjectSet*

Table A.5: Auxiliary functional modules in CRAFT.

Module	**Description**	**Input Types**	**Output Type**
`Unique`	Returns the single object from the input list, if the list has multiple elements returns `INVALID`	*ObjectSet*	*Object*
`Intersect`	Applies the set intersection operation	*(ObjectSet, ObjectSet)*	*ObjectSet*
`IntersectList`	Intersects an object set with multiple object sets	*(ObjectSetList, ObjectSet)*	*ObjectSetList*
`Difference`	Applies the set difference operation	*(ObjectSet, ObjectSet)*	*ObjectSet*
`ExistList`	Applies the Exist operation to each item in the input list returning a boolean list	*ObjectSetList / EventSetList*	*BoolList*
`AsList`	Returns an object set containing a single element specified by the input object	*Object*	*ObjectSet*

Table A.6: Event filter functional modules in CRAFT.

Module	**Description**	**Input Types**	**Output Type**
`FilterEvents`	Returns the list of events about a specific object from an event set	*(EventSet, Object)*	*EventSet*
`FilterCollision`	Returns the list of collision events from an event set	*EventSet*	*EventSet*
`FilterCollisionWithDynamics`	Returns the list of collision events involving dynamic objects	*EventSet*	*EventSet*
`FilterCollideGround`	Returns the list of collision events involving the ground	*EventSet*	*EventSet*
`FilterCollideGroundList`	Returns the list of collision event sets involving the ground	*EventSetList*	*EventSetList*
`FilterCollideBasket`	Returns the list of collision events involving the basket	*EventSet*	*EventSet*
`FilterCollideBasketList`	Returns the list of collision event sets involving the basket	*EventSetList*	*EventSetList*
`FilterEnterBasket`	Returns the In Basket events	*EventSet*	*EventSet*
`FilterEnterBasketList`	Returns the list of In Basket event sets	*EventSetList*	*EventSetList*
`FilterBefore`	Returns the events from the input list that happens before input event	*(EventSet, Event)*	*EventSet*
`FilterAfter`	Returns the events from the input list that happened after input event	*(EventSet, Event)*	*EventSet*
`FilterFirst`	Returns the first event	*EventSet*	*Event*
`FilterLast`	Returns the last event	*EventSet*	*Event*
`EventPartner`	Returns the object interacting with the input object through the specified event	*(Event, Object)*	*Object*
`FilterObjectsFromEvents`	Returns the objects from the specified events	*EventSet*	*ObjectSet*
`FilterObjectsFromEventsList`	Returns the list of object sets from a list of event sets	*EventSetList*	*ObjectSetList*
`GetCounterfactEvents`	Returns the event list if a specific object is removed from the scene	*Object*	*EventSet*
`GetCounterfactEventsList`	Returns the counterfactual event list for all objects in an object set	*ObjectSet*	*EventSetList*

A.2 Example Programs

Here we provide same functional programs for some of the questions given in Figure 1, which are used to extract the correct answers using our simulation environment. Figures A.3 to A.5 illustrate functional program samples that are designed for CRAFT Cause, Counterfactual, Descriptive, Enable and Prevent questions, respectively.

Question: *"Does the small brown sphere cause the tiny yellow box to enter the basket?"*

```
Var AffectorObject = FilterShape ( FilterColor ( FilterSize ( SceneAtStart(), "Small" ) , "Brown"), ‘‘Circle’’ )
Var PatientObject = FilterShape ( FilterColor ( FilterSize ( SceneAtStart(), "Small" ) , "Yellow"), "Cube" )
Exist (
        FilterStationary (
                Intersect (
                        Difference (
                                FilterObjectsFromEvents (
                                        FilterEnterBasket (
                                                Events()
                                        )
                                ),
                                FilterObjectsFromEvents (
                                        FilterEnterBasket (
                                                GetCounterfactEvents (
                                                        AffectorObject
                                                )
                                        )
                                )
                        ),
                        AsList ( PatientObject )
                ),
                StartSceneStep()
        )
)
```

Figure A.1: **Program for a sample Cause question from CRAFT.**

Question: *"How many objects fall to the ground if the small yellow box is removed?"*

```
Var QueryObject = FilterShape ( FilterColor ( FilterSize ( SceneAtStart(), "Small" ) , "Yellow"), "Cube" )
Count (
        FilterObjectsFromEvents (
                FilterCollideGround (
                        GetCounterfactEvents ( QueryObject )
                )
        )
)
```

Question: *"Will the small gray box enter the basket if any of the other objects are removed?"*

```
Var QueryObject = FilterShape ( FilterColor ( FilterSize ( SceneAtStart(), "Small" ) , "Gray"), "Cube" )
Var OtherDynamicObjects = Difference ( FilterDynamic ( SceneAtStart() ), AsList ( QueryObject ) )
AnyTrue (
        ExistList (
                IntersectList (
                        FilterObjectsFromEventsList (
                                FilterEnterBasketList (
                                        GetCounterfactEventsList ( OtherDynamicObjects )
                                )
                        ),
                        AsList (
                                QueryObject
                        )
                )
        )
)
```

Figure A.2: **Programs for two sample Counterfactual questions from CRAFT.**

Question: *"How many objects fall to the ground?"*

```
Count (
      FilterDynamic (
           FilterObjectsFromEvents (
                FilterCollideGround (
                     Events ()
                )
           )
      )
)
```

Question: *"After entering the basket, does the small yellow square collide with other objects?"*

```
Var QueryObject = FilterShape ( FilterColor ( FilterSize ( SceneAtStart(), "Small" ) , "Yellow"), "Cube" )
Var SmallYellowCubeEvents = FilterEvents ( Events(), QueryObject )
Exist (
        FilterAfter (
                FilterCollisionWithDynamics ( SmallYellowCubeEvents ),
                        FilterFirst (
                                FilterEnterBasket ( SmallYellowCubeEvents )
                        )
                )
        )
)
```

Figure A.3: **Programs for two sample Descriptive questions from CRAFT.**

Question: *"How many objects does the small gray block enable to enter the basket?"*

```
Var AffectorObject = FilterShape ( FilterColor ( FilterSize ( SceneAtStart(), "Small" ) , "Gray"), "Cube" )
Count (
        FilterMoving (
                Difference (
                        Difference (
                                FilterObjectsFromEvents (
                                        FilterEnterBasket (
                                                Events()
                                        )
                                ),
                                FilterObjectsFromEvents (
                                        FilterEnterBasket (
                                                GetCounterfactEvents (
                                                        AffectorObject
                                                )
                                        )
                                )
                        ),
                        AsList ( AffectorObject )
                ),
                StartSceneStep()
        )
)
```

Figure A.4: **Program for a sample Enable question from CRAFT.**

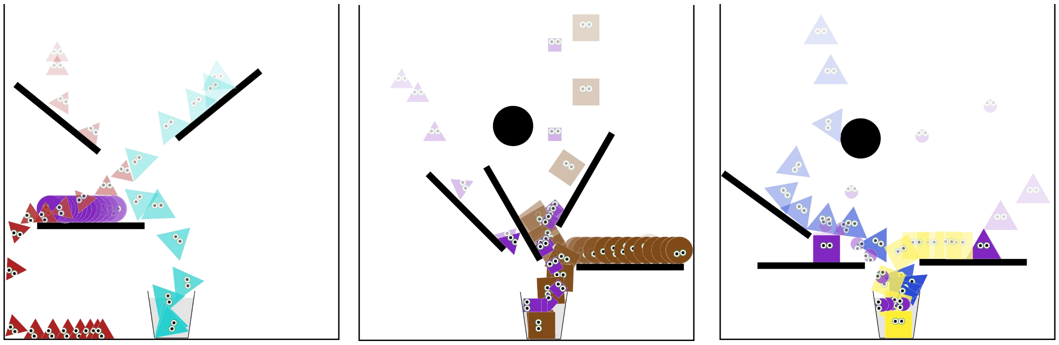
Question: *"Does the small yellow square prevent the tiny brown circle from entering the basket?"*

```
Var AffectorObject = FilterShape ( FilterColor ( FilterSize ( SceneAtStart(), "Small" ) , "Yellow"), "Cube" )
Var PatientObject = FilterShape ( FilterColor ( FilterSize ( SceneAtStart(), "Small" ) , "Brown"), "Circle" )
Exist (
        FilterMoving (
                Intersect (
                        Difference (
                                FilterObjectsFromEvents (
                                        FilterEnterBasket (
                                                GetCounterfactEvents (
                                                        AffectorObject
                                                )
                                        )
                                ),
                                FilterObjectsFromEvents (
                                        FilterEnterBasket (
                                                Events()
                                        )
                                )
                        ),
                        AsList ( PatientObject )
                ),
                StartSceneStep()
        )
)
```

Figure A.5: **Program for a sample Prevent question from CRAFT.**

A.3 Sample Predictions

In the main text, we only provide quantitative results. Here, in Figures A.6 to A.10, we include qualitative results showing the predictions of MAC-V, LSTM-CNN-V and MAC-F models, which are found to be the three top-performing models, on some sample Cause, Counterfactual, Descriptive, Enable and Prevent questions from CRAFT, respectively.



There is a large purple circle, does it trigger the big cyan triangle to go into the bucket?

Correct Answer: “True”

MAC-V: “False”
LSTM-CNN-V: “False”
MAC-F: “False”

How many objects does the big brown block stimulate to go into the container?

Correct Answer: “1”

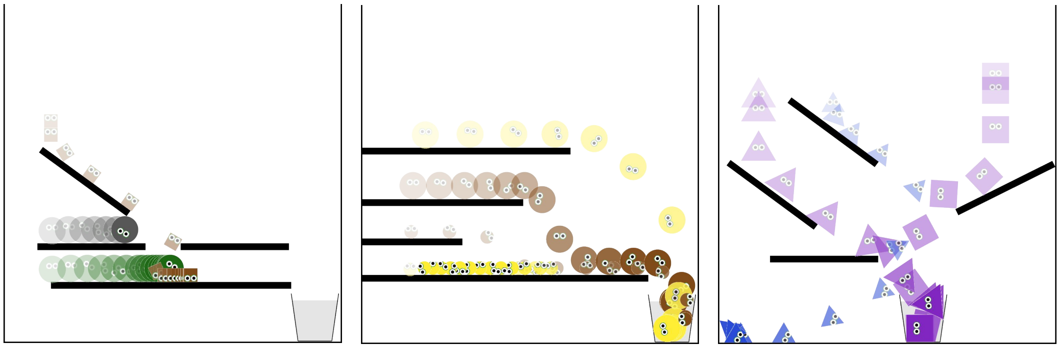
MAC-V: “1”
LSTM-CNN-V: “1”
MAC-F: “1”

How many objects does the big purple triangle trigger to hit the floor?

Correct Answer: “0”

MAC-V: “1”
LSTM-CNN-V: “1”
MAC-F: “1”

Figure A.6: **Sample Cause questions from CRAFT and the predictions obtained with MAC-V, LSTM-CNN-V, and MAC-F models.**



If the large green circle is removed, how many objects fall into the bucket?

Correct Answer: “0”

MAC-V: “0”
LSTM-CNN-V: “0”
MAC-F: “2”

If the big brown circle is removed, how many objects end up in the basket?

Correct Answer: “3”

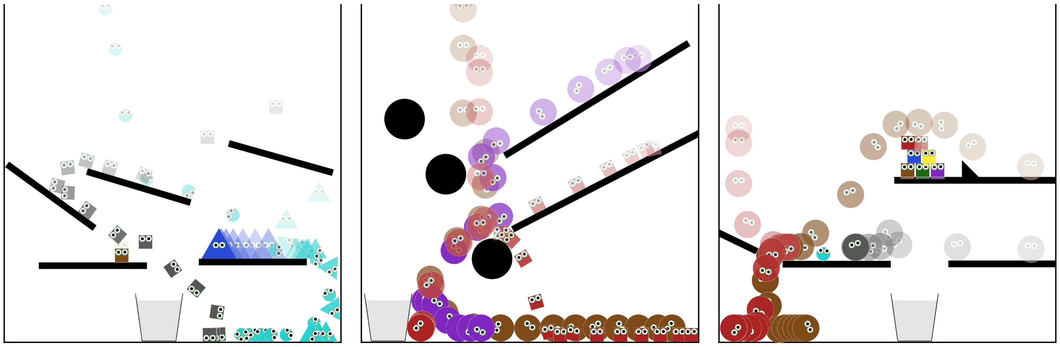
MAC-V: “2”
LSTM-CNN-V: “2”
MAC-F: “2”

If any of the other objects are removed, will the small blue triangle get into the container?

Correct Answer: “False”

MAC-V: “True”
LSTM-CNN-V: “True”
MAC-F: “False”

Figure A.7: **Sample Counterfactual questions from CRAFT and the predictions obtained with MAC-V, LSTM-CNN-V, and MAC-F models.**



How many objects fall to the ground?

Correct Answer: “3”

MAC-V: “0”
LSTM-CNN-V: “2”
MAC-F: “0”

How many objects hit the ground?

Correct Answer: “4”

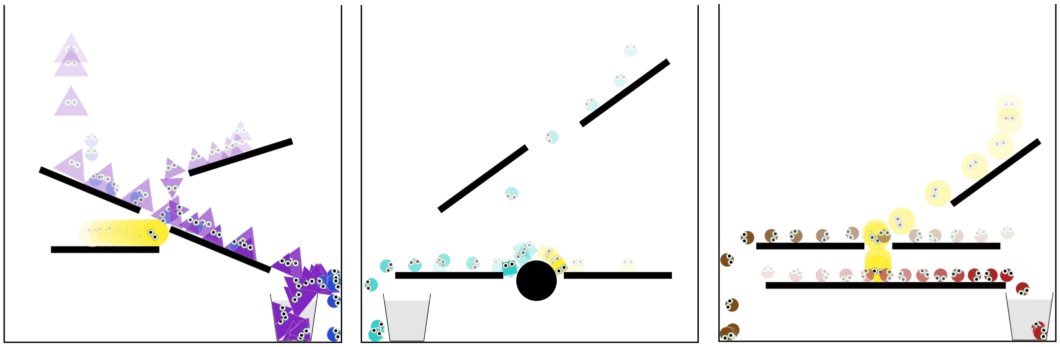
MAC-V: “2”
LSTM-CNN-V: “3”
MAC-F: “2”

What color is the object the large brown circle first collides with?

Correct Answer: “Red”

MAC-V: “Brown”
LSTM-CNN-V: “Cyan”
MAC-F: “Yellow”

Figure A.8: **Sample Descriptive questions from CRAFT and the predictions obtained with MAC-V, LSTM-CNN-V, and MAC-F models.**



Does the big purple triangle enable the small blue circle to hit the floor?

Correct Answer: “False”

MAC-V: “True”
LSTM-CNN-V: “True”
MAC-F: “False”

Does small yellow cube allow the tiny cyan circle to hit the floor?

Correct Answer: “False”

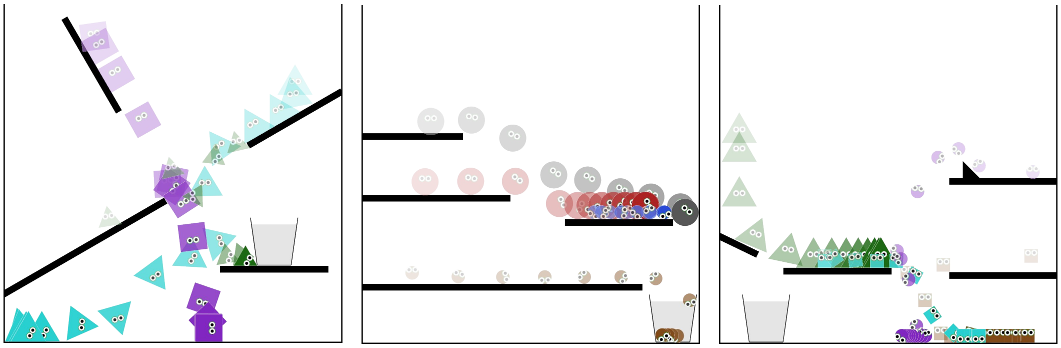
MAC-V: “False”
LSTM-CNN-V: “False”
MAC-F: “False”

There is a large yellow circle, does it help the tiny red circle to end up in the basket?

Correct Answer: “False”

MAC-V: “True”
LSTM-CNN-V: “False”
MAC-F: “False”

Figure A.9: **Sample Enable questions from CRAFT and the predictions obtained with MAC-V, LSTM-CNN-V, and MAC-F models.**



How many objects are blocked by the tiny green triangle from hitting the ground?

Correct Answer: “0”

MAC-V: “0”
LSTM-CNN-V: “0”
MAC-F: “1”

Does the big red circle hold the tiny blue circle from going into the bucket?

Correct Answer: “True”

MAC-V: “False”
LSTM-CNN-V: “False”
MAC-F: “False”

Is the big green triangle blocked by the tiny cyan cube from falling to the ground?

Correct Answer: “True”

MAC-V: “False”
LSTM-CNN-V: “False”
MAC-F: “False”

Figure A.10: **Sample Prevent questions from CRAFT and the predictions obtained with MAC-V, LSTM-CNN-V, and MAC-F models.**

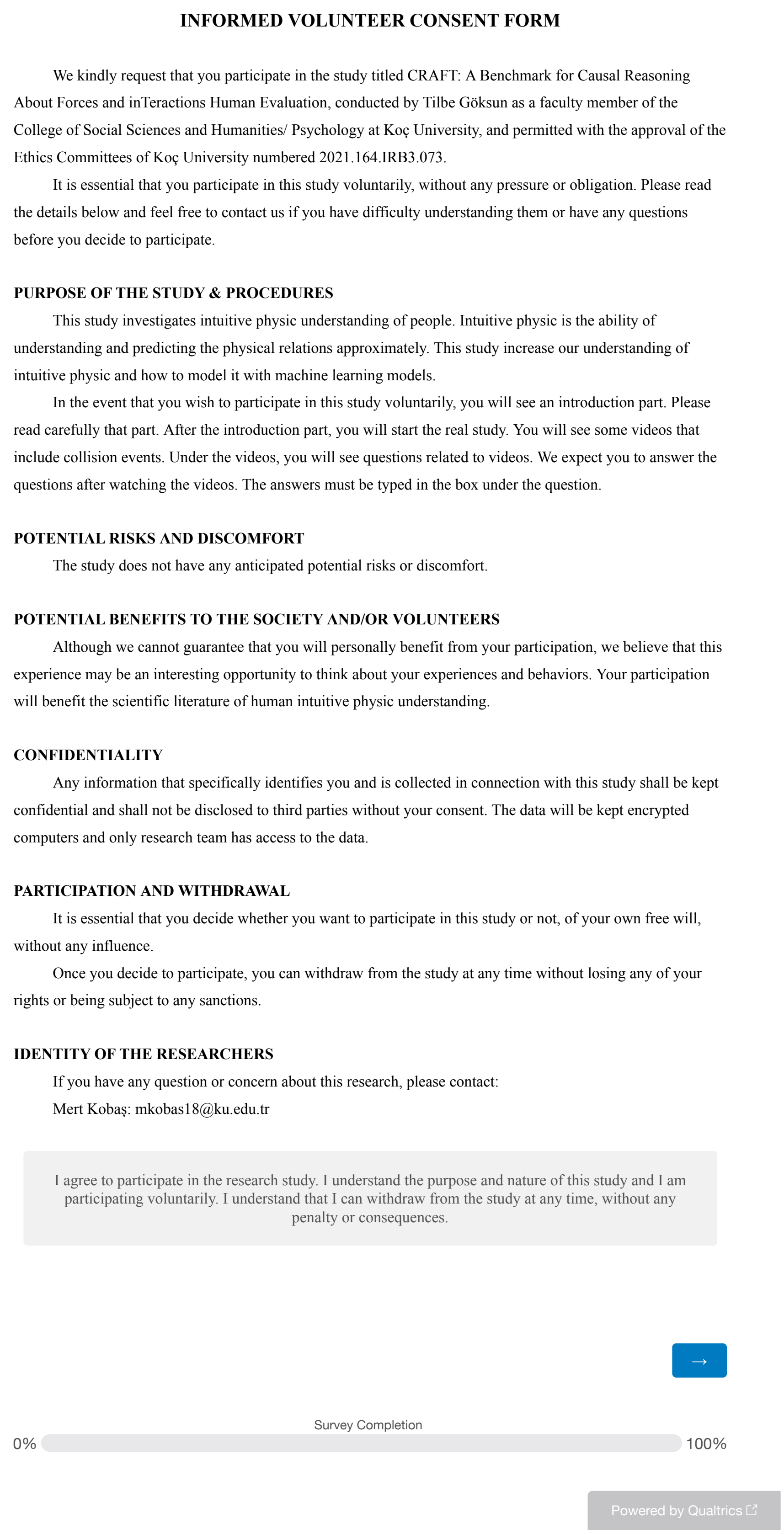
A.4 Human Evaluation

The data from human participants were collected online via Qualtrics. The approximate time to complete the study was between 20 and 30 minutes. Participants did not take any bonus or wage. They attended the study voluntarily. The personal identifying information was not obtained. There were not expected negative outcomes of the study on participants, but they could leave the study whenever they want. Koç University’s Institutional Review Board approved the study (Protocol no: 2021.164.IRB3.073). The consent form can be found in Figure A.11.

For the human evaluation, the participants saw the videos and multiple choice questions all together. The instruction that was given to participants is shown below:

> In this study, you will be asked to answer questions related to the videos that include interactions between some moving or stationary objects. For example, two objects might collide with each other, one may enter the basket or hit to the ground. The questions will be about:
>
> - Counting the number of objects took place in a certain event (consider only dynamic objects unless stated otherwise). Example: “How many objects enter the container?”
> - Whether an object help/hinder a specific event. Example: “There is a big green block, does it allow the small blue circle to enter the basket?”
> - Imagining what would happen if a certain event occurs. Example: “If any of the other objects are removed, will the small yellow triangle go into the bucket?”
> - Questioning the shape/color of an object. Example: “What color is the object the tiny brown triangle last collides with?”
> - We ask you to watch each video first and then answer the question related to the video later. You can re-watch each video until you move to the question related to the video. For the yes/no questions, you are only allowed to select “yes” or “no”. Descriptive questions relating to the number of objects should be answered with sliding the bar. When you are ready, you can click “Next” to start answering the next question.

The instruction page can be found in Figure A.12.



INFORMED VOLUNTEER CONSENT FORM

We kindly request that you participate in the study titled CRAFT: A Benchmark for Causal Reasoning About Forces and inTeractions Human Evaluation, conducted by Tilbe Göksun as a faculty member of the College of Social Sciences and Humanities/ Psychology at Koç University, and permitted with the approval of the Ethics Committees of Koç University numbered 2021.164.IRB3.073.

It is essential that you participate in this study voluntarily, without any pressure or obligation. Please read the details below and feel free to contact us if you have difficulty understanding them or have any questions before you decide to participate.

PURPOSE OF THE STUDY & PROCEDURES

This study investigates intuitive physic understanding of people. Intuitive physic is the ability of understanding and predicting the physical relations approximately. This study increase our understanding of intuitive physic and how to model it with machine learning models.

In the event that you wish to participate in this study voluntarily, you will see an introduction part. Please read carefully that part. After the introduction part, you will start the real study. You will see some videos that include collision events. Under the videos, you will see questions related to videos. We expect you to answer the questions after watching the videos. The answers must be typed in the box under the question.

POTENTIAL RISKS AND DISCOMFORT

The study does not have any anticipated potential risks or discomfort.

POTENTIAL BENEFITS TO THE SOCIETY AND/OR VOLUNTEERS

Although we cannot guarantee that you will personally benefit from your participation, we believe that this experience may be an interesting opportunity to think about your experiences and behaviors. Your participation will benefit the scientific literature of human intuitive physic understanding.

CONFIDENTIALITY

Any information that specifically identifies you and is collected in connection with this study shall be kept confidential and shall not be disclosed to third parties without your consent. The data will be kept encrypted computers and only research team has access to the data.

PARTICIPATION AND WITHDRAWAL

It is essential that you decide whether you want to participate in this study or not, of your own free will, without any influence.

Once you decide to participate, you can withdraw from the study at any time without losing any of your rights or being subject to any sanctions.

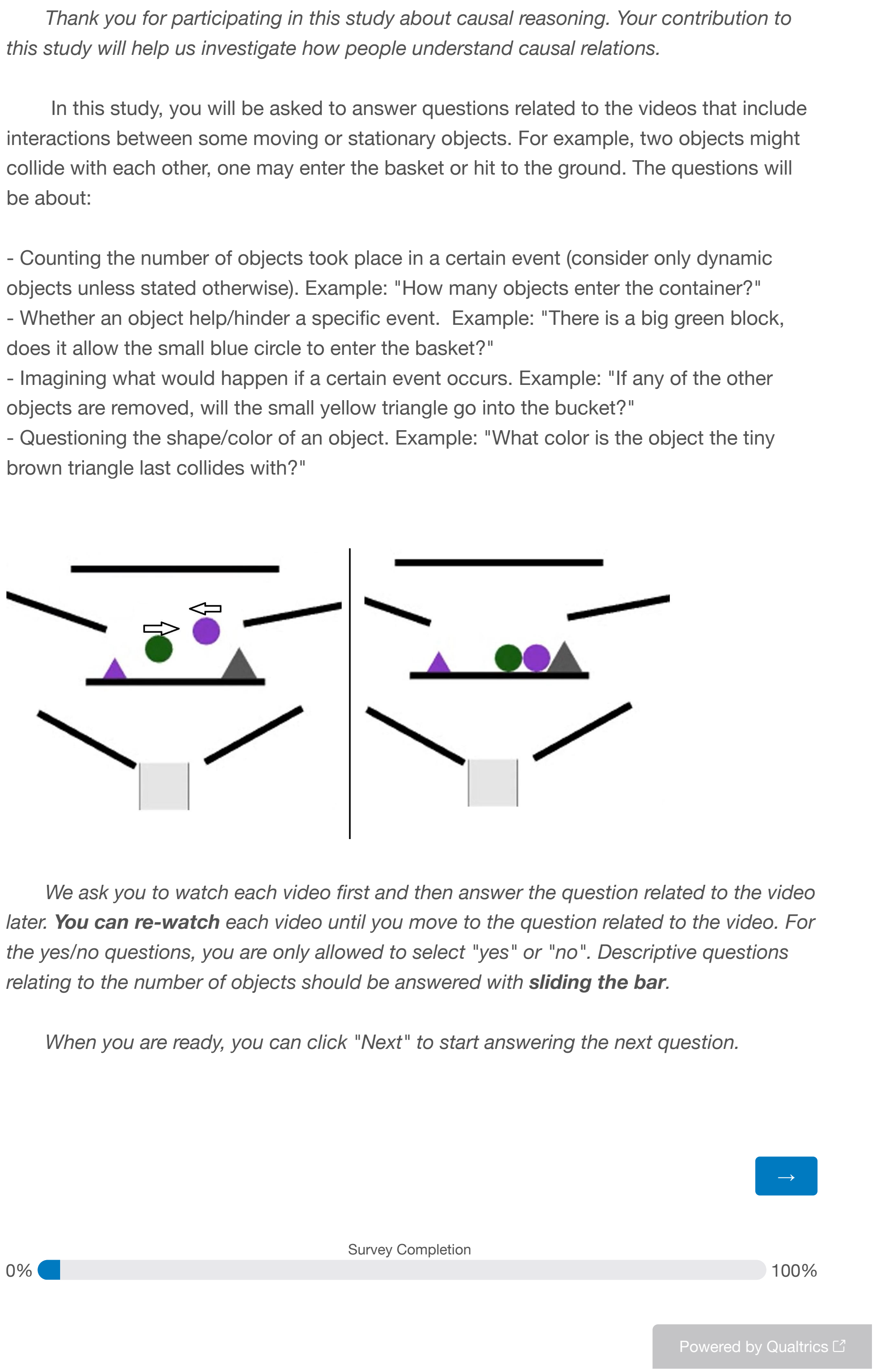
IDENTITY OF THE RESEARCHERS

If you have any question or concern about this research, please contact:

Mert Kobaş: mkobas18@ku.edu.tr

I agree to participate in the research study. I understand the purpose and nature of this study and I am participating voluntarily. I understand that I can withdraw from the study at any time, without any penalty or consequences.

Figure A.11: **The obtained consent form that describes potential risks and IRB approvals.**



Thank you for participating in this study about causal reasoning. Your contribution to this study will help us investigate how people understand causal relations.

In this study, you will be asked to answer questions related to the videos that include interactions between some moving or stationary objects. For example, two objects might collide with each other, one may enter the basket or hit to the ground. The questions will be about:

- Counting the number of objects took place in a certain event (consider only dynamic objects unless stated otherwise). Example: "How many objects enter the container?"
- Whether an object help/hinder a specific event. Example: "There is a big green block, does it allow the small blue circle to enter the basket?"
- Imagining what would happen if a certain event occurs. Example: "If any of the other objects are removed, will the small yellow triangle go into the bucket?"
- Questioning the shape/color of an object. Example: "What color is the object the tiny brown triangle last collides with?"

We ask you to watch each video first and then answer the question related to the video later. ***You can re-watch*** *each video until you move to the question related to the video. For the yes/no questions, you are only allowed to select "yes" or "no". Descriptive questions relating to the number of objects should be answered with* ***sliding the bar****.*

When you are ready, you can click "Next" to start answering the next question.

Figure A.12: **The information form of the human evaluation study.**

B Datasheet for CRAFT

This document is prepared in accordance with the guideline suggested in Datasheets for Datasets [**?**]. The most updated version can be found here.

Motivation

For what purpose was the dataset created?

CRAFT was created in order to facilitate research on understanding and closing the gap between the capabilities of human intelligence and artificial systems in grasping and reasoning about physical relationships between different objects in an environment.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset was created by Tayfun Ates, M. Samil Atesoglu, Cagatay Yigit, Erkut Erdem from Hacettepe University and Ilker Kesen, Mert Kobas, Aykut Erdem, Tilbe Goksun and Deniz Yuret from Koç University.

Who funded the creation of the dataset?

CRAFT was supported in part by GEBIP 2018 Award of the Turkish Academy of Sciences to E. Erdem and T. Goksun, BAGEP 2021 Award of the Science Academy to A. Erdem, and AI Fellowship to Ilker Kesen provided by the KUIS AI Center.

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?

The instances of CRAFT include a video, a question about the video, its answer, the functional program which is the ground-truth process that is used to answer the question, the states of dynamic objects and static scene elements at the start of the simulation and at the end of the simulations, causal graph of the events occurred in the video, variation videos which are created removing each dynamic object one by one, and lastly the states of objects and causal graphs for variation videos.

How many instances are there in total (of each type, if appropriate)?

CRAFT contains 58K video and question pairs that are generated from 10K videos from 20 different virtual environments.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

Please refer to Section 3 of the main paper for a detailed description of the sampling procedure used to generate questions.

What data does each instance consist of?

The video and question-answer pairs are used as the basic components for this visual question answering study. The question about the video is asked to an artificial model or a human subject. The test containing multimodal inputs question the capabilities of the subject in understanding and reasoning about physical relationships occurring in an environment. We use other instances in the dataset to find answers to questions automatically and share them for further analysis if required. Functional programs can run on object states and causal graphs to find the answer. Moreover, they can be integrated in training process for different models as well. Similarly, if ground-truth information regarding object states and causal graphs can also be extracted. Furthermore, some questions require counterfactual analysis that we define using variation videos formally. In order to evaluate effect of an object on the scene, we remove it an re-simulate the environment. We share instances regarding variations for further analysis.

Is there a label or target associated with each instance? If so, please provide a description.

Each instance consists of a ground-truth answer associated with the question about a dynamic scene.

Is any information missing from individual instances? We do not provide object-level segmentation maps.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?

Instances are generated from 20 different scene layouts with some randomization.

Are there recommended data splits (e.g., training, development/validation, testing)?

We share CRAFT with two different split alternatives that we call easy and hard settings. Both of the alternatives contain non-overlapping train, validation, and test set. There are 20 distinct layouts from which we created our virtual scenes for CRAFT. In easy setting, each split might contain images from all of the scene layouts. On the other hand, in hard setting, train, validation, and test splits contain images from 12, 4, and 4 of the 20 layouts, respectively. That is, in the hard setting, the corresponding test samples are generated from unseen scene layouts.

Are there any errors, sources of noise, or redundancies in the dataset?

The process that we followed to make sure that the answers are not affected much with the slight perturbations to the initial states is described in Section 3 of the main paper.

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

The dataset is self-contained.

Does the dataset 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.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

No.

Does the dataset relate to people?

No.

Does the dataset identify any subpopulations (e.g., by age, gender)?

No.

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.

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.

Collection Process

How was the data associated with each instance acquired?

All instances of CRAFT are generated automatically using a physics engine.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?

We use Box2D physics simulator [**?**] to create our visual scenes, extract object states and causal graphs. Furthermore, we extend the work CLEVR [**?**] to create CRAFT questions and answers.

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

The dataset is generated from scratch and it does not depend on an already existing dataset.

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)?

Authors prepared the scripts which create visual and textual data automatically.

Over what time-frame was the data collected?

Data generation scripts ran about 51 hours to create 9917 videos and 57524 questions.

Does the dataset contain all possible instances?

Although we provide all instances for this version of CRAFT, it is possible for anyone to create new samples by running the scripts provided in our code repository.

If the dataset is a sample, then what is the population?

Please refer to Section 3 of the main paper for a detailed description of the sampling procedure used to generate questions.

It is possible the enlarge CRAFT by running existing scripts to obtain huge amount of data because of the randomness existing in video generation process as described in the paper. New dynamic objects, static scene elements, events can also be created to enrich CRAFT. Moreover, it is also possible to add new types of scene layouts and question categories or types. For example, CRAFT focuses on mostly physical reasoning. It is possible to add tasks questioning different capabilities of Humans such as spatial reasoning, planning, and so on. There is actually no limit for creating datasets similar to CRAFT.

Were any ethical review processes conducted (e.g., by an institutional review board)?

Koç University's Institutional Review Board approved the user study (Protocol No: 5152021.164.IRB3.073).

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

The data from human participants for the user study were collected online via Qualtrics.

Were the individuals in question notified about the data collection? Yes.

Did the individuals in question consent to the collection and use of their data? The participants of the user study are asked to sign the consent form given in Figure A.8.

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?

Not applicable.

Preprocessing/Cleaning/Labeling

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)?

There were two preprocessing steps applied to the dataset. Firstly, after creating a video and question-answer pair, we applied simple perturbations by changing certain values of dynamic objects slightly at the start of the simulation and re-simulated the video. If the answer to the question is changed in any of the variations, then we removed the video and the question pair from the dataset. Secondly, in order to obtain a dataset which is uniform as possible in all dimensions, we removed video and question pairs whose answers are dominant after the first perturbation filter.

By collecting this dataset, we had the chance to observe that although the artificial systems have demonstrated incredible progress in the past decade, there are still areas that should be investigated for them. Therefore, CRAFT can be considered as a sample dataset which will facilitate the research in closing the gap between humans and artificial systems.

Preprocessing steps achieve two main aims of ours. Firstly, we wanted to eliminate video and question pairs whose answers are inconsistent between different variations of the same video with small perturbations. We observed that these were the cases for which humans subjects had some troubles. Secondly, we wanted to make CRAFT difficult enough for machine reasoning models by aiming at avoiding learning shortcuts by selecting the most frequent answers in answering questions. The second step of preprocessing procedure mostly achieves this aim.

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

The raw data were saved, but were not made public.

Is the software used to preprocess/clean/label the instances available?

We plan to publicly release the software used to generate the scenes and the questions.

Distribution

Has the dataset been used for any tasks already?

We have used the dataset to train unimodal and multimodal baselines described in the paper.

Is there a repository that links to any or all papers or systems that use the dataset?

Links to the related papers will be listed in the project website at `https://sites.google.com/view/craft-benchmark`.

What (other) tasks could the dataset be used for?

Since the sample videos in our dataset include interactions between the objects themselves and the environment, they can be used in problems such as future state prediction and video generation.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?

No.

Are there tasks for which the dataset should not be used?

No.

Uses

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?

CRAFT is publicly available at `http://github.com/hucvl/craft/`.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?

The dataset is available through our project website and GitHub. Large dataset files are stored on Zenodo.

When will the dataset be distributed?

The dataset was first released in June 2021.

What license (if any) is it distributed under?

The dataset is released under MIT license.

Maintenance

Who is supporting/hosting/maintaining the dataset?

CRAFT will be supported and maintained by the authors, M. Samil Atesoglu and Cagatay Yigit.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Email contact: `matesoglu@hacettepe.edu.tr`, `cyigit@hacettepe.edu.tr`.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?

Extending CRAFT in different directions is planned. All versions of CRAFT will be available at `http://github.com/hucvl/craft/`.