Supplementary Material for CLEVRER-Humans: Describing Physical and Causal Events the Human Way

Jiayuan Mao*
MIT

Xuelin Yang*
Stanford University

Xikun Zhang
Stanford University

Noah D. Goodman
Stanford University

Jiajun Wu
Stanford University

We bear all responsibility in case of violation of rights. The data created or used during this study are openly available on the project's website (https://sites.google.com/stanford.edu/clevrer-humans). We confirm that the data is under CC0 license. We will provide maintenance to the website and dataset regularly and upon request.

The rest of this supplementary document is organized as the following. First, in Section A we provide visualizations of data collected in CLEVRER-Humans, as well as more analysis on the comparison between human causal judgments and heuristics-based labels. In Section B, we describe the implementation details of models studied in the main paper and add additional failure case analysis of models. Next, in Section C, we describe the user interface for dataset collection. Finally, in Section D, we supplement dataset sheets for CLEVRER-Humans.

A Dataset Visualization and Analysis

Fig. 2 and Fig. 3 show the example graph collected in the stage I (causal event cloze) and stage II (binary CEG labelling), respectively. First, Fig. 2 shows that the causal cloze tasks can progressively collect a large number of human-written event descriptions by re-using the response of previous annotators. On average, we can obtain 29.4 descriptions per video, highlighting the advantage of our design. Second, the condensed CEGs contains high-quality causal relations of physical events, as shown in Fig. 3. It demonstrates both the language diversity and the richness of causal relations in the CEGs of CLEVRER-Humans. These figures provide a straightforward illustration of our data collection pipeline and the quality of our data.

A.1 Dataset Statistics

First, CLEVRER-Humans contains dense annotations of causal relations between physical events. Fig. 1a and Fig. 1b show the distributions of the number of nodes and edges in each CEG. The average number of CEG nodes is 4.71 and the average number of labeled edges is 12.7. These dense annotations of CEGs form the rich and complicated causal structures in our dataset.

Second, CLEVRER-Humans offers diverse free-form language descriptions while retaining balances in object properties. Fig. 1c shows the length distribution of event descriptions: the average length is 7.00 (as a reference, the average event description length of CLEVRER is 8.93). CLEVRER-Humans has a vocabulary length of size 219, which is much greater than CLEVRER (82). Fig. 1d and Fig. 1e

*indicates equal contribution. Correspondence to: jiayuanm@mit.edu.

Submitted to the 36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks. Do not distribute.

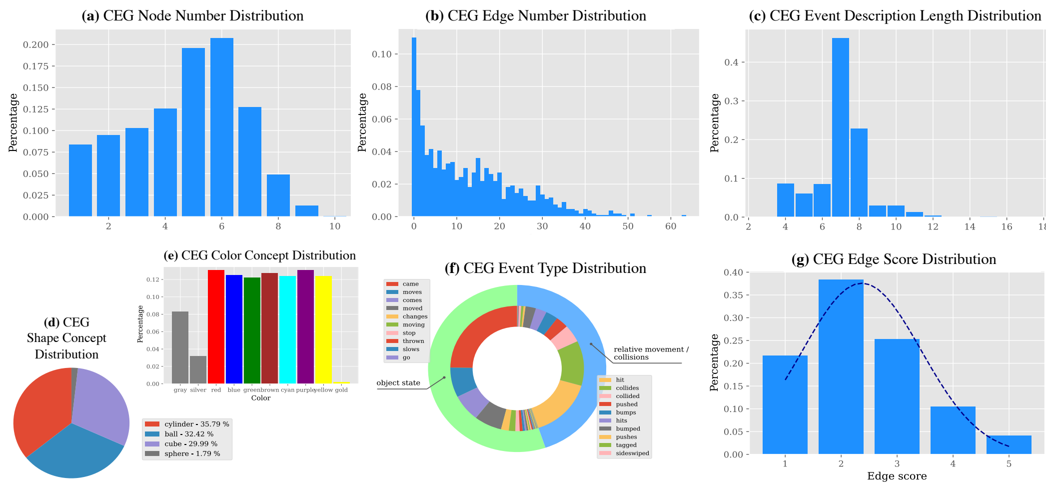


Figure 1: Statistics on the CLEVRER-Humans dataset. From left to right, the first row figures are distributions of (a) the number of nodes per CEG, (b) the number of edges per CEG, and (c) sentence lengths excluding the "which of the following is responsible for" prefix. The second row figures are distributions of (d) object shapes, (e) colors, (f) event type attributions based on verbs, and (g) CEG edge scores labelled by MTurkers, respectively.

show the distribution of object property concepts: colors and shapes. They remain unbiased when considering the synonyms such as “ball” and “sphere” and “gray” and “silver.”

Next, most importantly, CLEVRER-Humans engage a variety of physical events for causal reasoning tasks. In particular, Fig. 1f shows the distribution of event types computed based on the main verb of the event description. The outer circle represents the general event families. The corresponding inner breakdowns display more than 10 variations of the expression based on verbs for each event type. In comparison, the original CLEVRER dataset contains only three event types (and verbs): enter, exit, and collide. Therefore, CLEVRER-Humans significantly improves the diversity and brings in a challenge for machines to recognize and ground these events in practice.

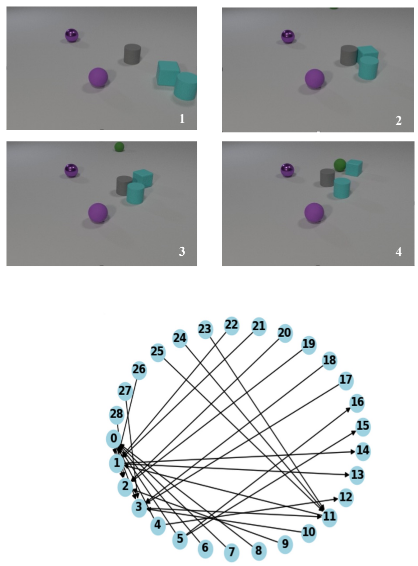
In the following box, we list all verbs that have been annotated by human annotators and generated by our machine generative model. We have lemmatized all verbs to remove the tense.

> come, move, change, stop, throw, slow, go, travel, begin, spin, roll, stand, halt, roll, lose, leave, head, want, hurl, enter, hit, collide, push, bump, push, tag, sideswipe, bounce, strike, touch, cause

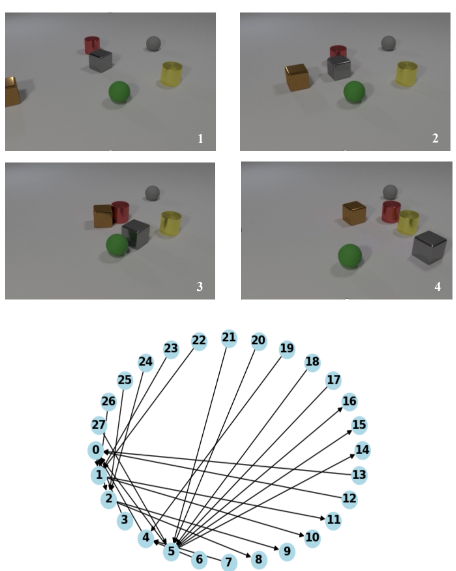
We also would like to point out that for some verbs, if they seem to be synonyms (e.g., bump and sideswipe), they can have subtle differences in physical grounding. For example, A bumps into B usually implies that A is moving faster than B and its collision changed the state of B. Furthermore, different tense of the same verb have different meanings in sentences: "the event that ball A moved is responsible for the collision" is different from "the event that ball A is moving is responsible for the collision." In the former case, ball A does not have to be moving while the collision happens.

It is possible to hand-craft a lot of rules to handle each individual cases (e.g., bump, sideswipe, roll), but that will require additional hyperparameters for thresholds, and may be hard to align with human perception.

Finally, CLEVRER-Humans’ annotation reflects the subjective judgment of causality in physical events. CLEVRER-Humans offers 5 choices when asking MTurkers to label the causality level. Fig. 1g shows the distribution of edge scores with an average of 2.37. Note that this distribution is skewed towards lower scores (as shown by the Gaussian approximation in the dotted curve). This reflects the fact that most event pairs do not have causal relationships. Finally, although we have



0: the cyan cube sideswipes the grey cylinder
1: the green ball hits the cyan cube
2: the green ball touches the grey cylinder
3: the green ball bumps into the cyan cube
4: the cyan cube was moving in direction opposite to the grey cylinder
5: the cyan cylinder came from below
6: the cyan cylinder struck the cyan cube
7: the cyan cube was struck by the cyan cylinder
8: the cyan cylinder moved
9: the green ball was pushed by the cyan cube
10: the gray cylinder moved
11: the green ball came from above
12: the square cyan was hit by the cyan cube causing the grey cylinder to move away
13: the cyan cube is pushed to the right side
14: the cyan cube spins counterclockwise
15: the green ball collided with the cyan cube
16: grey cylinder moved forward
17: the blue rubber cylinder bumped the cyan cube in the way of the green ball
18: the green ball slides down into the scene into the cyan cube
19: the green ball touched the cyan cube
20: the green ball bounces off the cyan cube and touches the grey cylinder
21: cyan is on a direct path to green ball path
22: the cyan cylinder hits the cyan cube
23: the green ball was already moving downwards
24: the green ball had velocity
25: the green ball rolled down into the screen and bounced into the cyan cube
26: the cyan cylinder hit the cyan cube
27: the green ball and the cyan cube were heading to the same point
28: the green ball bounces off the blue square which was traveling in the opposite direction



0: the brown cube starts to spin clockwise
1: the brown cube hits the silver cube and then hit the red cube
2: the brown cube came from the west
3: the brown cube starts to spin clockwise because gold metal shere to swipe
4: the brown cube hits the metal purple cube
5: the brown cube hit the ash cube by the side
6: the metal purple cube was in the way
7: the red cylinder collides with the metal purple cube
8: the grey cube touched the green ball
9: the grey cube moved south
10: the red cube ends up in the east
11: the silver cube hits the green ball
12: the brown cube collided with the silver cube after the silver cube was struck by the red cylinder
13: the brown cube collided with the grey cube
14: the ash cube moves to the right
15: the gray cube spun
16: the silver cube hit the green ball
17: the red cylinder collided with the yellow cylinder
18: the red cylinder pushed the ash cube away from the brown cube's path
19: the red cylinder hit the metal purple cube
20: the brown cube came from the left
21: the red cylinder hit the ash cube from the top
22: ricochets off to left and bounces into red cube
23: the brown cube came from the side
24: the brown cube started off on the west and was moving east
25: the brown cube was pushed by an unknown and unseen force
26: the silver cube absorbed the force from the brown cube in order to hit the red cube
27: the red cylinder hit the grey cylinder which changed its direction

Figure 2: Visualization of two samples of annotations collected in Stage I causal cloze tasks. They are collected progressively by feeding the response of a user as the input of another one. The black arrows indicate the annotation orders.

binarized the edge labels for the sake of consistency with CLEVRER, the raw score-based judgment can be potentially helpful in other tasks such as cognitive science studies.

Therefore, we can conclude that CLEVRER-Humans is a high-quality causal relation dataset with significantly more diverse event types and language descriptions than CLEVRER.

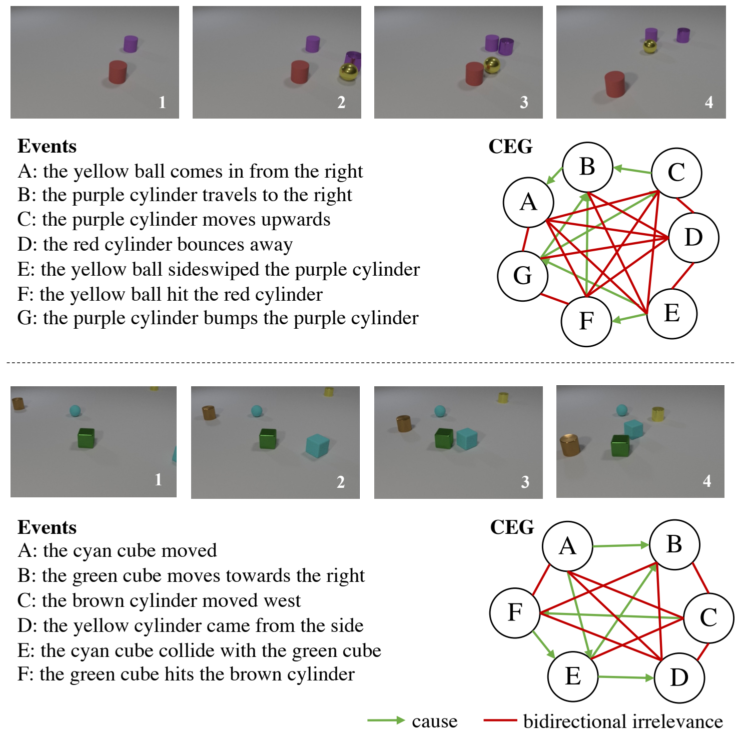


Figure 3: Visualization of two samples of CEGs in CLEVRER-Humans. The green arrows represent causal relations and the red edges represent bidirectional irrelevance. We can see the rich causal relations among physical events presented in the CEGs.

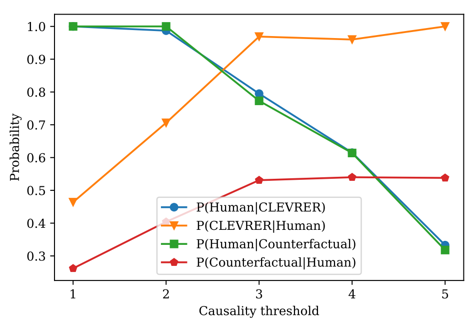


Figure 4: Effect of different causality thresholds on the binarized human causal relation. The x-axis is the ablation threshold (i.e., 4 means a score $\geq$ 4 represents a causal relation). The y-axis is the conditional probability.

	Y = y_1	Y = y_2	Y = $y_1 \wedge y_2$	Y = $y_1 \vee y_2$	Y = $y_1 \oplus y_2$
P(X = Human | Y)	0.62	0.61	0.23	0.34	0.34
P(Y | X = Human)	0.96	0.54	0.29	0.62	0.33

Table 1: Comparison between different combinations of heuristics-generated causal labels and human labels, on a sampled subset of CLEVRER [1]. The entry P(X|Y) denotes the fraction of event relations that are annotated as causal by protocol X given that the relations are annotated as causal by protocol Y. y_1, y_2 denote the existence of causal relations defined CLEVRER's heuristic and Counterfactual causal relation, respectively.

A.2 Comparison between Heuristic and Human Causal Judgments

We supplement the effect of different thresholds on the graded causal relation in Fig. 4. In the human performance study, we asked the participants to choose a threshold from 1-5 if they had to binarize their judgment. The average threshold suggested by the participants is 3.6. In practice, we choose a threshold of 4 to obtain the causal relation that humans are more certain about.

Having shown the two common heuristics-generated causal labels (CLEVRER's and counterfactual intervention) diverges from human judgment, we also provide the results on comparisons between different combinations of heuristics-generated causal labels and human judgments. We use the logic operators and ($\wedge$), or ($\vee$), xor ($\oplus$). As shown in Table 1, none of these combinations can give an close enough approximation to human judgment, which further justifies our motivation to use human labeled causal data for CLEVRER-Humans.

B Implementation Details

In this section, we present the implementation details of our neural-network-based event description generator, the baseline models studied in the main paper, and the error bars for models across different random seeds.

B.1 Stage II Implementation

We first describe the input pre-processing for neural event generators. For each object, we concatenate the one-hot encoding of physical properties (including shapes, colors, and materials) and the motion information (including location, orientation, velocity, angular velocity, and whether the object is inside the camera view) in each of the 128 frames in a video. For each object, at each time step, the input dimension is 24.

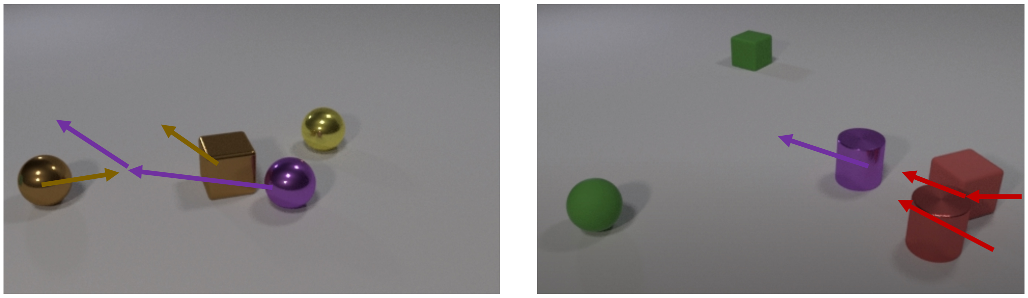
Our rule-based event detector for object pairs works as the following. For object pairs, we first extract all segments that are composed of consecutive frames when two objects are close two each other. Specifically, we say two objects are close if the L_∞ norm of the displacement vector between two objects is smaller than 0.5 meter (i.e., their x, y, z displacements are all smaller than 0.5 meter). Within each segment, the event detector predicts event types including moving together, object approaching, and collision, based on changes in the motion information. For example, if two objects are physically close for more than 20 frames without rapid changes in velocity, we consider them relatively static, thus "moving together." If both objects change directions within their close period, we consider a collision happened. We can further distinguish the changes in relative positions (either "bouncing back" or "one approaching another") by the sign of the dot product of velocity vectors. For any object pair, if no events are detected in the course of the entire video, we do not include this pair for future captioning.

After getting the input sequences, we use neural event generators consisting of an encoder and a decoder to produce captions. The encoder uses a linear layer and a GRU unit to encode the input sequence [2]. The decoder applies Softmax on the embedding of input and the hidden state to produce the attention weights. It then uses GRU and a linear layer to produce an English caption of specific objects in the video. Single-object and pairwise captioning models share the same architecture but are trained independently. The hidden dimension of both the encoder and the decoder is 256 for single-object models and 128 for pairwise models. The dropout rate for the decoder is set to 0.3 for single-object models and 0.1 for pairwise models. All models are trained with a learning rate of 0.001. For the grammar check module in the post processor, we drop the sentences with two consecutively repeated words. We also exclude the sentences that miss verbs or verb arguments, such as sentences ending with words "from," "to," "at," "is," etc.

B.2 Baseline Implementation

Language-Only models. For the language-only models, we use a LSTM [3] with GloVe [4] word embedding. The hidden dimension of 512 and the dropout rate is 0.2. We use the Adam optimizer [5] with a learning rate is 4×10^{-4} and a weight decay of 10^{-5}. The batch size is 4. Following the data splits in CLEVRER, we split 20% of the training pairs as the validation set and choose the model with the highest validation accuracy.

CNN+LSTM. For the CNN+LSTM models, we use a pre-trained ResNet-50 to extract 2,048-dimensional features from the video frames [6]. We uniformly sample 25 frames for each video as input. The word embedding for questions is initialized by the GloVe [4] word embedding. Both LSTMs (the question encoder and the video sequence encoder) have 1 layer with a hidden dimension of 512. We apply a dropout rate of 0.2 on the input layer and 0.5 on the hidden layers. We use the Adam [5] optimizer with a weight decay of 5×10^{-4}. The learning rate is 10^{-5} training from scratch and 10^{-3} for finetuning. The batch size is 128 for both trained-from-scratch and finetuning experiments. We split 20% of the training pairs as the validation set and choose the model with the highest validation accuracy.



Question: What is responsible the purple ball collides with the brown cube?
Choice: The purple ball comes up.
Answer: Wrong
Model prediction: Correct

Question: What is responsible for the red cube sideswiped the purple cylinder?
Choice: The red cube bumped the red cylinder.
Answer: Wrong
Model prediction: Correct

Figure 5: Examples of common prediction errors. The arrows in the image represent the moving direction of objects of interest in the video. **Left:** failure caused by nuances in human language. While the purple ball is constantly moving upwards coordinate-wise, humans understand the phrase "comes up" as more of the later part of the trajectory (after the purple ball collides with the brown ball). Therefore, machines cannot give a correct prediction. **Right:** failure in bridging the domain shift. Humans may consider the change in the trajectory to be minor and appears not to be a deciding factor of the outcome event, but the model predicts it as a cause following similar heuristics in CLEVRER.

BERT+LSTM We supplement CNN+BERT models as a model with a stronger text encoder. The CNN is the same as in the CNN+LSTM baseline. We use the pretrained BERT uncased base model from HuggingFace library [7]. The BERT tokenizer is set to max length 32 padding and truncation. During training, We fix weights of the text encoder. We use the Adam optimizer with a weight decay of 5×10^{-4}, and a learning rate of 10^{-5} training from scratch and 10^{-3} for finetuning. The batch size is 128. We choose the model with highest validation accuracy with 20% of the training set as the validation set.

ALOE. We implement our model based on the publicly released code [8]. Since the public release does not contain training code, we implement the training procedure using the following settings. For object embeddings, we use the pre-trained MONet embeddings released by the authors. For optimization, we use the Adam [5] optimizer with a weight decay of 10^{-3} (we have also benchmarked 10^{-2}, 10^{-3} and 10^{-4}). We split 5% of the training pairs as the validation set and choose the model with the highest validation accuracy.

B.3 Error analysis

We summarize the common failures of the models: for pretrained-only models, the common error comes from the failure of incorporating more diverse language and events. For example, as shown in Table 2, the program parser of NS-DR and VRDP fails to generate proper programs for descriptions in CLEVRER-Humans. The deficiency of language understanding often leads to wrong predictions.

For training from scratch models, one possible reason to the test errors is the nuances in human language. Specifically, models do not only need to identify the objects being referred to but also their physical properties: the cause of "the red cube slows down" can be hard to identify because speed does not appear to be as explicit as other properties such as colors and shapes. As shown in our comparison between human judgement and heuristics-based causal judgements, the nuance in language can influence human judgments, posing difficulties for machines to ground the events and simulate the reasoning process. For instance, the left figure in Fig. 5 illustrates the nuances in language resulting a discrepancy between human judgment and prediction. Moreover, for large

Event	Parsed program
The purple sphere slows down from the right.	["events", "objects", "purple", "filter_color", "sphere", "filter_shape", "unique", "filter_collision", "objects", "unique", "filter_color", "sphere", "filter_shape", "unique", "filter_collision", "unique"]
The red ball comes to a stop.	["events", "objects", "red", "filter_color", "unique", "filter_collision", "objects", "red", "filter_color", "unique", "filter_collision", "unique"]
The yellow cube comes from the right side at a fast speed.	["events", "objects", "yellow", "filter_color", "cube", "filter_shape", "unique", "filter_out", "unique"]

Table 2: Examples of errors produced by program parser. In the first row, the model cannot identify the event "slow down from the right" and gives incorrect parsing to find another object involved in a collision ("filter_collision"). In the second row, the model cannot represent the event "come to a stop" due to the expand in vocabulary and gives an incorrect output ("filter_collision"). In the third row, the model mistakenly represents the enter event as the exit event ("filter_out") because the description is more complicated in CLEVRER-Humans. We follow the notation of programs as in NSDR and VRDP.

models such as ALOE, learning to simulate human reasoning process from scratch based on very little data can be difficult, especially with limited training size.

For finetuned models, we have not seen significant improvement brought by the pretraining phase. This is primarily because of the domain gap between human judgement and heuristics-based labelling. Specifically, our human experiments have shown that p(Human | CLEVRER-Heuristic) = 0.62. That is, only 62% of the event pairs that have been labelled as causal in CLEVRER, are labelled as causal by human annotators. The right figure in Fig. 5 gives an example of the error caused by the domain shift. Future work may consider other ways of pretraining, such as pretraining on event recognition, which may be more transferable, and pretraining with other types of heuristics.

C Labeling Interface

We develop labeling interfaces based on boto3 with Amazon MTurk python API. We include example trials of both the causal cloze tasks and CEG annotation tasks in Fig. 6 and Fig. 7, respectively. The full instruction texts are provided on the labeling page of our project's website. The estimated hourly pay to the Mechanical Turk participants is about \$6.1 and the total amount spent on participant compensation is about \$3500. Specifically, the cloze tests and part of the pairwise causal relationship annotations were completed by users from the U.S., and the pay was \$7.7/hour (above federal minimum wage). At a later stage of our project, we were unfortunately constrained by the budget available to us and opened the tasks to workers outside the U.S. Thus, overall, our average hourly wage is \$6.1. Our goal has always been to commit to best practice and offer fair pay to users whenever possible, and we will continue to do so in the future.

In causal cloze tasks, the participants are asked to write an event description given a cause or outcome event, as shown in Fig. 6. We specify the expectation of responses (such as using complete sentences, avoiding ambiguous third-person pronouns, etc.) in the instruction. We design a small comprehension quiz with 7 multiple choice questions and 2 chances to submit to ensure the participants understand the instructions correctly.

In the CEG annotation tasks, the participants are asked to label the correctness and causal relation of two event descriptions as shown in Fig. 7. We give 4 examples with detailed explanations to help them understand the rationale of the task. We also give an illustration of object colors and shapes for reference. Bounding boxes are added to the videos to accelerate the process of locating objects of interest.

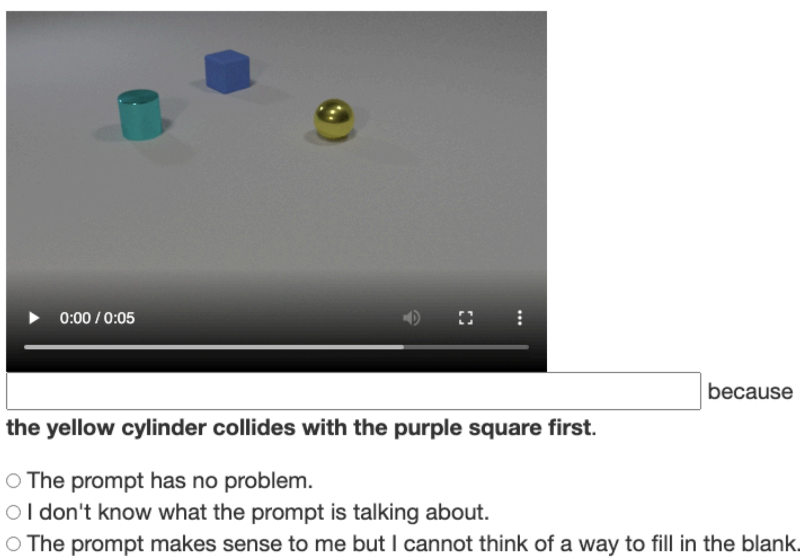


Figure 6: Example of a causal cloze trial. The participants are asked to fill in the blank after watching a video. They can also select the checkboxes if they do not understand the prompt.

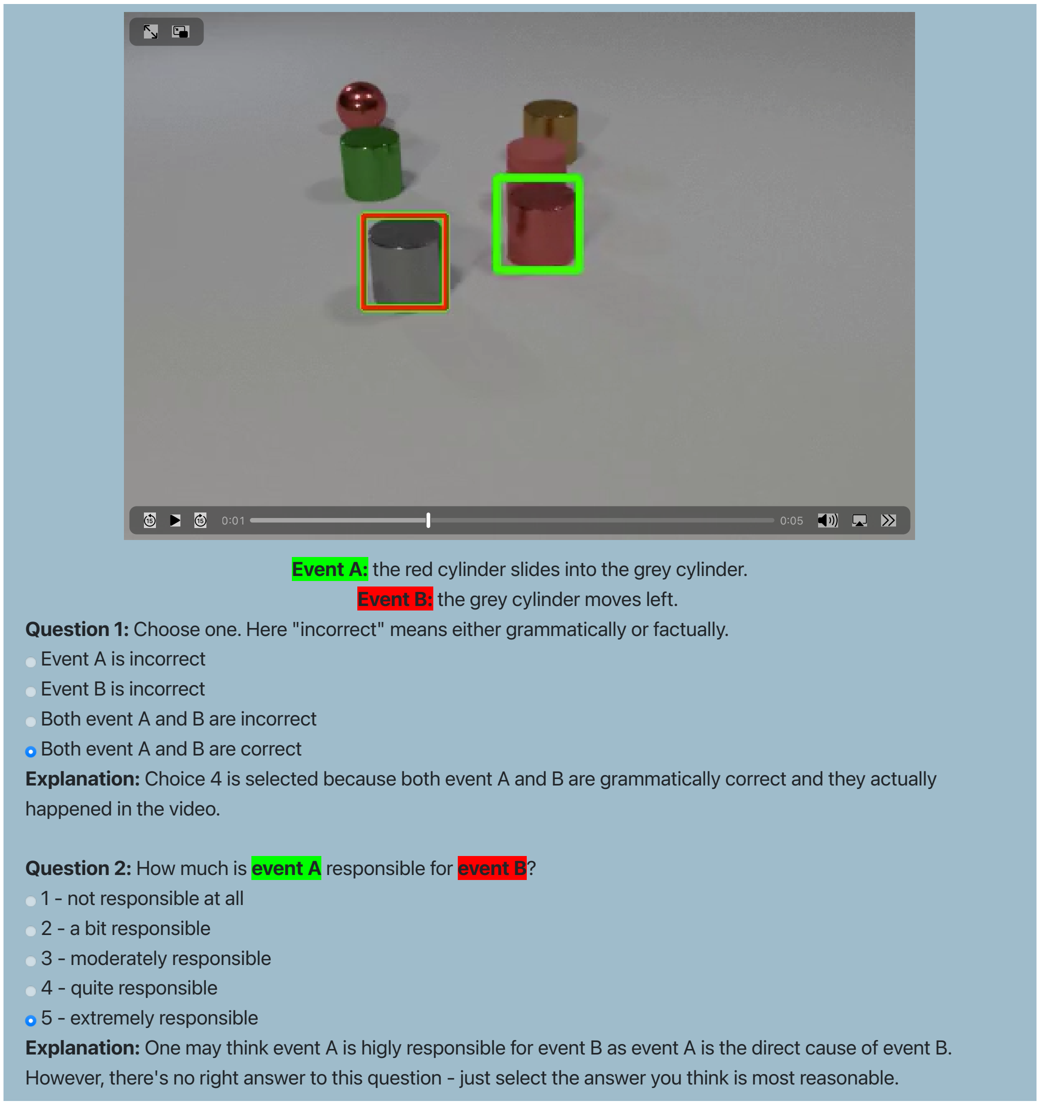


Figure 7: Example of a CEG trial. The participants are first asked to select if the event descriptions are correct. If both correct, they are asked to label the level of causal relations between the descriptions. For each event pair, we provide bounding boxes for objects involved in the events for better annotation efficiency.

D Dataset Release

Our dataset is under CC0 License. We provide a documentation using data statements for NLP in [9].

Short form data statement CLEVRER-Humans is a large-scale video reasoning dataset of human-annotated physical event descriptions and their causal relations. It contains machine-generated texts based on crowdsourcing data in US English (en-US). The language quality and causal structure annotations are obtained by watching videos, reading texts, and entering responses on MTurk.

The following is the long form data statement of CLEVRER-Humans:

Curation Nationale CLEVRER-Humans contains descriptions and causal relations of physical events such as an object entering the scene and two objects colliding with each other. The goals in selecting texts were to ensure the interpretability and correctness of the descriptions and to provide a variety of free-form captioning of physical events in videos. We first collected human written event descriptions by causal cloze tasks, then used machine learning models to generate more natural language descriptions based on the curated data. We post-processed the data by grammar checking, object existence checking, and verb re-balancing. Finally, we obtained human annotations on the texts through crowdsourcing: if the labelers annotated the texts interpretable and correct, we ask them to provide a pairwise graded causal judgment of the events.

Language Variety The event description data for causal cloze tasks were collected on MTurk. Information about which varieties of English are represented is not available, but at least CLEVRER-Humans includes US (en-US) mainstream English.

Speaker Demographic We used a cascaded generator composed of a rule-based event detector and a neural pairwise generator to generate texts. When the curating training data in causal cloze tasks, we restricted the location of these MTurkers to be in the US. It is expected that most speakers use English as their native language. Estimated demographics of MTurkers may refer to [10].

Annotator Demographic We hire the MTurkers with the approved HITs of 1000 or higher. We expect the MTurkers to be the general public who are familiar with basic crowdsourcing process. When collecting data, we release the tasks in batches, where each HIT contains 30 QA pairs mostly coming from one or two videos. We perform quality check to unsure annotators have sufficient knowledge of English language. We also answer their questions about the annotation process by emails. It is expected that most speakers use English as their native language. Estimated demographics of annotators may refer to [10].

Speech Situation The intended audience of the texts is the general public. The texts are all in written form. MTurkers are expected to read the text and watch the video when doing causal cloze and causal labeling tasks. The video is about 5 seconds, which can be played as many times as one wishes.

Text Characteristics The texts are plain English descriptions of a physical event in a video. A sentence usually contains one or more physical object(s) (i.e. sphere, cylinder, or cube) and the related movements or interactions presented in the video. Ideally, the generated event descriptions can maintain the vocabulary and structural characteristics similar to the training data from causal cloze tasks. The detailed statistics of the text are shown in the Dataset Statistics section.

Recording Quality N/A

Other N/A

Provenance Appendix The videos of CLEVRER-Humans are from the CLEVRER dataset [1].

D.1 Intended Use

CLEVRER-Humans can be used as a benchmark in physical scene understanding and causal reasoning. It evaluates machines ability to understand and analyze physical interventions in a restricted setting. Machines are provided with a short video and expected to answer questions regarding the causes of events in the video.

D.2 Maintenance Plan

We will host our dataset permanently on our project's website. Users are granted access to the dataset through links on the website. We provide versioning of the dataset and archive backup regularly.

D.3 Quality Check

Quality checks over CEG node correctness are performed by majority voting. Since we have split the annotation of each video to 3 annotators, and they will see overlapping events and annotate their correctness. Checks for edge correctness are performed by including additional "quality checking" questions. Specifically, each annotator will see 3 videos and 10 questions for each video. 1 of the video will be from a small and manually-curated dataset by authors, containing 30 videos. The entire answer set will be accepted if and only if the annotators answers those quality-checking questions correctly (more specifically, have a small divergence with our answer).

References

[1] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. CLEVRER: Collision events for video representation and reasoning. In *ICLR*, 2020. 4, 11

[2] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv:1409.1259*, 2014. 6

[3] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 6

[4] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014. 6

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6, 7

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 7

[8] David Ding, Felix Hill, Adam Santoro, Malcolm Reynolds, and Matt Botvinick. Attention over learned object embeddings enables complex visual reasoning. https://github.com/deepmind/deepmind-research/tree/a5522d078413e340a2caa1ab0c86d76d2b7efa40/object_attention_for_reasoning, 2021. 7

[9] Emily M. Bender and Batya Friedman. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *TACL*, 6:587–604, 12 2018. 11

[10] Joel Ross, Lilly Irani, M. Six Silberman, Andrew Zaldivar, and Bill Tomlinson. Who are the crowdworkers? shifting demographics in mechanical turk. In *CHI Extended Abstracts on Human Factors in Computing Systems*, 2010. 11