Supplementary Material for STAR: A Benchmark for Situated Reasoning in Real-World Videos

1 Broader Impacts

The objective of our work is exploring human-like reasoning in daily life situations. Thus the situated reasoning benchmark in videos can demonstrate the merit of situated reasoning to human cognition, as opposed to the alternative focused more on data-driven perception. It helps motivate more research leveraging knowledge logic across disciplines, in this case, connecting cognitive science and machine learning. If successful, the system can help develop robotics comprehension or interaction and intelligent information triage system in real-world settings. Like other AI systems, the potential negative impact will be the loss of human job opportunities displaced by the automatic agents. The consequence caused by systems failures could be false predictions due to the complexity of the given scenarios. All video data (relating to persons) in our dataset are from the public dataset Charades [10]. They prevent privacy issues by data anonymization with anonymous IDs. For details, please refer to their homepage. We will further explore other dimensions in the future.

2 Dataset Statistics

Data Type With hypergraph representation of situations, we can construct a controlled and compositional data space through sampling entities, action hyperedges, and corresponding data types. In total, we select 29 entities (28 objects and one person entity), 111 transitive action predicates, and 24 relationships. All objects have multiple connections with actions (action verb) or persons in situations. By default, our situations are only contained a single person. Action predicates are compositions of an atomic action verb and an object noun (*e.g.*, take the book, sit on the chair *etc*..). Meanwhile, STAR has the following relationships in situations: person-object contracting relationships and spatial relationships, object-object spatial relationships, or action-action temporal relationships. We collect about 570K relations in total by marrying labeled human-object relationships and generated static scene graphs. We start from the detected relationships of [2] and expand relationships by using scene graph generator [11].

Vocabulary Distribution In Section 3.1 of the main paper, we introduced how to construct a controllable data type space for real-world dynamic situations in STAR. The vocabulary of the cleaned and processed data types are shown in Table 1. The composition of rich verbs and objects generates diverse actions of STAR; interactive, spatial, and temporal relationships associate the entities or situations in situation videos.

In Figure 1 (a), the word cloud (word size indicates frequency) represents popular objects, verbs, and relationships in STAR. We can see that the benchmark has diverse daily-life entity nouns or action verbs, and the words with the highest frequency are "`take`", "`put`", and "`cup`". STAR actions are consist of action verbs and objects. In Figure 1 (b) and (c) analyze the diversity of the combination between verbs and objects. We only adopt reasonable actions which are *compositional* verb-object pairs, like "`take book`" and "`close door`". Inversely, STAR avoids to use unreasonable verb-object pairs like "`eat table`", "`take floor`". We select those verb-object pairs in which each verb has at least two corresponding objects and vice versa. This removes the verbs or objects which only have one combination and increases the complexity of question answering. It also stops the models from inferring the corresponding object or verb by simple collocations.

35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks.

Scenes
Hallway , Bedroom, Home Office, Closet, Kitchen, Basement, Recreation room, Entryway, Living room, Dining room, Laundry room
Objects
broom, picture, phone/camera, floor, blanket, closet/cabinet, pillow, bed, sofa/couch, box, blanket, cup/glass/bottle, shoe, clothes, door, window, shelf, sandwich, mirror, laptop, table, clothes, food, bag, dish, refrigerator, paper/notebook, medicine, laptop, towel
Verbs
open, close, hold, put, take, throw, grasp, wash, pour, drink, walk, stand, turn, eat, run, sit, lie, watch, tidy
Relationships
eating, wiping, touching, standing on, writing on, drinking from, wearing, holding, carrying, twisting, covered by, sitting on, leaning on
on, behind, in front of, at, under, near, beneath, on the side of, above, over
before, after

Table 1: The vocabulary of the benchmark STAR.

Datas Split STAR is split into training, validation, and testing sets with a ratio of about 6:1:1 (8 partitions in total). The data partition is carefully designed to ensure no video overlap among subsets, which selects the unseen videos in the validation or testing sets. The setting avoids the model overfitting for the seen videos in the training set.

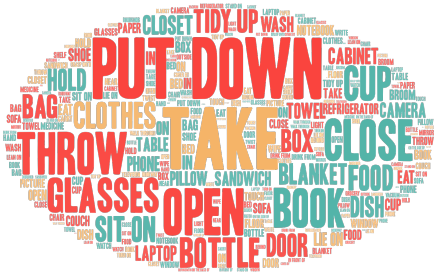
Question or Program Complexity Analysis STAR benchmark generates questions and answers automatically. The complexity of the questions and answers reflect the benchmark reasoning difficulty. Hence, we provide complexity analysis for questions and programs in Figure 2. We group questions or programs by their length or logic steps. Meanwhile, we group questions by their question type to see how they are distributed on the STAR.

The questions that contain less than 10 words are “simple" questions. Questions with more than 10 but less than 13 words are treated as “medium" questions. Questions containing more than 13 words are named “hard" questions. The question length in STAR is carefully balanced. Noted that such sentence length only reflects the complexity of question text but not the reasoning complexity.

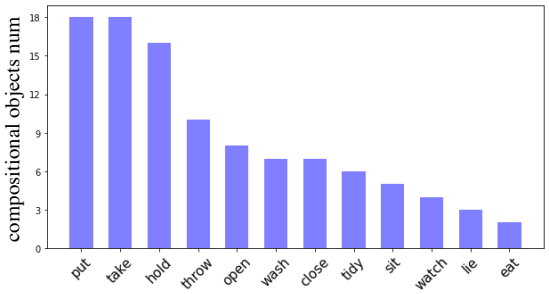
The number of program steps represent the answering logic complexity. We define programs that have less than 10 steps as “simple" programs. Those with more than 10 but less than 13 steps are defined as “common“ programs. And the programs are regarded as “complex" if they have more than 13 steps. As shown in Figure 2, most of questiones in STAR contains a common logic that are neither too simple nor too complex. Table 5 and Table 6 illustrate data types and program functions.

3 Implementation Details

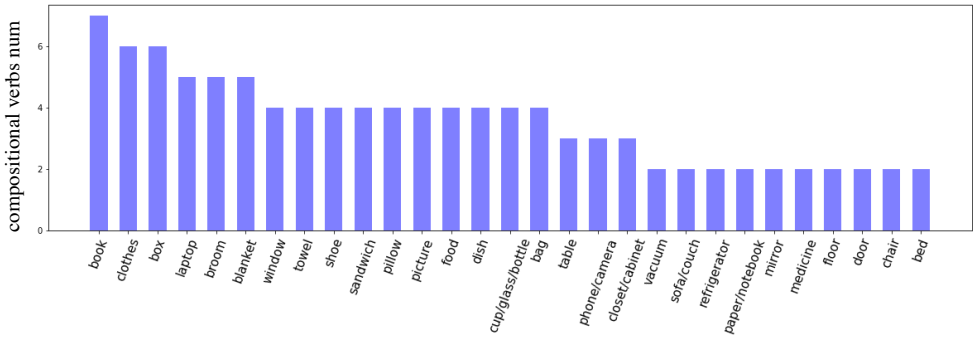
In our experimental comparison, all baselines use the same settings of visual and linguistic representations. All models use videos as inputs. In our model, we sample 16 keyframes for each situation video and resize keyframes to fixed size 224×224. In CNN-LSTM, we take 3D visual features extracted by a pretrianed ResNext101 [12] (on Kinetics 400 [4]). Our implemented CNN-BERT use the same 3D visual features to represent visual tokens. Notably, we freeze feature extraction models (like ResNet-50, ResNext 101) for feature extracting, and train all baseline reasoning models from scratch on STAR only which means there is no pretrained model. Beyond frame-level features, we also have Video QA baselines (*e.g.*, HRCN) which take the video motion features as visual input (*i.e.*, C3D). For the necessary baselines, the same object detector RestNext [12] is adopted to extract visual embedding (2048-dim) of entities in the situations. The video parser we used for detecting objects and their relationships is VCTree with TDE [11] pretrained on ImageNet [1] and fintune on STAR. Questions, answers, and options are represented by 300-dimensional pretrained



(a) Word clouds for frequent objects, verbs and relationships.



(b) The compositional objects statistics



(c) The compositional verbs statistics

Figure 1: The distribution of the vocabulary and the statistics on the compositonal pairs. Best viewed in color.

GolVe[8] word embeddings. We tuned hyperparameters (learning rate, dropout) independently per model, uses early stopping to avoid overtraining, and selected the best checkpoints based on the validation error. In our NS-SR, we train and test each question type separately with the learning rate is $5 * 10^{-5}$, and the batch size are 128, 64, 32, 32 for Interaction, Sequence, Prediction and Feasibility question respectively. We used Tesla V100 GPUs with batch size 16 on a single GPU to train the Transformer-based Action Transition Model with 6 layers of transformers. Both visual and language embeddings are projected to 256-dimension vectors and combined as the sequence input. Since the neural-symbolic model may get predictions that are out of candidate options while imperfect searching on the situation graphs. Once the predicted option is out of candidate options, the model would randomly select answers from the candidate options. Such random selections are under fixed random seed and repeat 10 times per question to ensure stability. We optimized all experiments on the validation set firstly, and then evaluated them on the test set, so they were generalized from the validation set to the test set.

4 Model Details

Situation Hypergraph Encoding As mentioned in Section 5.2 of the main paper, the NS-SR performs dynamic state transitions over situation hypergraphs for each question, and the transformer-based encoder is a significant submodule. To remain structured representations of situation hypergraphs, the NS-SR encoder transforms a set of situation hypergraphs into structured token sequences (called hypergraph token sequences) for each situation video. To indicate the graph entities, hyper-edges, segments, and situations, we sum the multiple types of embedding vectors for each token: 1) token embedding: object appearances, human poses, relationship classes or predicate classes, 2) type or hyperedge embedding: indicates action predicates, persons, objects or relationships 3) situation embedding: records situation time-order, 4) position embedding: object and person bounding boxes, and 5) segment embedding. We also use separation tokens to mark the boundary of situations and

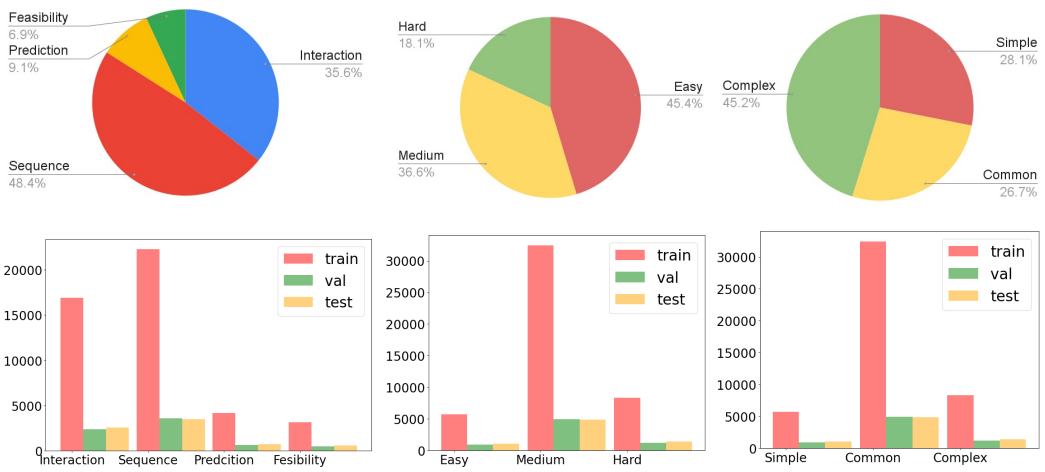


Figure 2: The complexity analysis for questions and programs. Pie Charts: question number distribution of question type, question length and logic steps of nested programs. Histograms: the distributions organized by data splitting. Best viewed in color.

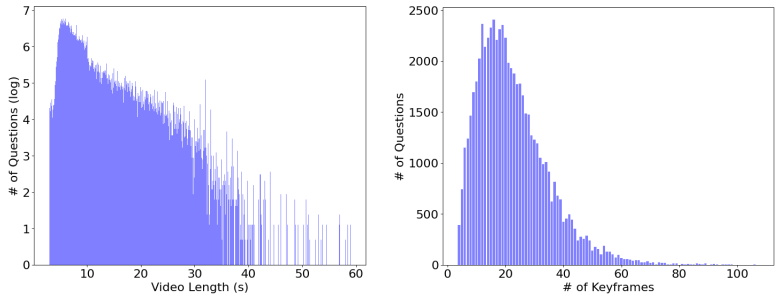


Figure 3: Left: the distribution of situation video length. Right: the distribution of keyframe amount for each situated question in STAR. Most of situated questions contains 8 to 30 situation subgraphs as well as the corresponding keyframes. Best viewed in color.

segments in token sequence. We adopt 2, 16 and 8 for max action number, max situation graph number and max relationships number respectively. Such parameters represent that we set there are maximum 16 situation graphs in a situated video, maximum 2 actions and maximum 8 relationships in a single situation graph. Overall, we demonstrate an example of a hypergraph token sequence with two and more hypergraph units in Figure 4. Each situation column is a structured situation sub-graph, and the associated situation subgraphs share the same action and are connected by action hyper-edges. Each row means a specific embedding token type.

5 Experiment Details

Experiment Settings of the Diagnostic Model Evaluation In Table 4, we provide the settings of all model variants in the diagnostic model evaluation.

input token sequence	Situation 1							Situation 2								Situation N
token emb (vector)	ACT_Emb	ACT_Emb	[SEP]	PER_Emb	~~REL_Emb~~	OBJ_Emb	[SEP]	PER_Emb	REL_Emb	~~OBJ_Emb~~	ACT_Emb	ACT_Emb	[SEP]	~~PER_Emb~~	~~REL_Emb~~	~~OBJ_Emb~~	[SEP]	...	[END]
type emb (discrete)	A	A	M	P	R	O	M	P	R	O	A	A	M	P	R	O	M	...	M
segment emb (int)	0	0	Max	1	1	1	Max	2	2	2	0	0	Max	1	1	1	Max	...	Max
hyper-edge emb (int)	1	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	...	End
situation emb (int)	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	...	N

Figure 4: The structure of hypergraph token sequence for representations in the encoding of the NS-SR. Best viewed in color.

Modules (Settings)	Implementations	Performances
Obj Detector (Obj Det)	Faster-RCNN [9] (X101-FPN)	mAP=25.67
Rel Detector (w/ Obj Det)	VCTree [11] (TDE-sum)	R@20=29.01, R@50=30.31
Rel Detector (w/ Obj GT)	VCTree [11] (TDE-sum)	R@20=40.34, R@50=41.07
Language Parser	Seq2Seq	Acc=99.80
Action Transition Model (w/ Obj Det and Rel Det)	our transformer	action Acc=25.21, Obj Acc=30.61, Rel Acc=35.42
Action Transition Model (w/ Obj GT and Rel Det)	our transformer	action Acc=40.73, Obj Acc=72.74, Rel Acc=49.69
Action Transition Model (w/ Obj GT and Rel GT)	our transformer	action Acc=51.88, Obj Acc=73.13, Rel Acc=66.45

Table 2: Quantitative Evaluation of Perception Modules on STAR. Metric Notations: Acc: class prediction accuracy, mAP: mean average precision at IOU=0.5, R@K: recall at top K.

Method	Visual Perception	Interaction	Sequence	Prediction	Feasibility
VisualBERT [7]	perfect Obj bbox GT	34.67	35.89	31.18	31.35
ClipBERT [6]	perfect Obj bbox GT	36.32	38.88	30.73	29.76

Table 3: Performances of NN Models with Perfect Perception on STAR

Quantitative Results of Perceptions To know "imperfect" perceptions quantitatively mentioned in Section 5.2, we conducted new experiments on the STAR test set for vision, language, and hypergraph models respectively, as shown in Table 2. For each model, we use the same metrics with the reported question accuracy for comparison.

For object or relationship detectors, we adopted pre-trained Faster-RCNN (X101-FPN)/ VCTree (TDE-sum) models and fine-tuned them on STAR training set annotations until the losses converged. For the variants with object ground-truth, we use bounding box annotations from [2] to localize the objects/humans in situation videos. Compare with the reported performances in the state-of-the-art scene graph models [11, 2], Table 2 shows our detectors achieve comparable or higher performances, although their detectors are trained on higher-quality (Visual Genome [5]) or larger-scale (Action Genome [2], 0.47M object or 1.7M relationship instances) data. Thus, our visual perception models are well-trained.

Visual perception is an important challenge on STAR besides the structured abstraction. The NS-SR model (or video parser module) inputs for visual perception are videos instead of hypergraphs; Action transition module (structured abstraction) inputs are built upon on video parser outputs (objects or relationships). Meanwhile, without perfect visual perceptions, the performances of variants decrease significantly with a 6% to 14% drop in average accuracy, as shown in the diagnostic evaluation. As an upstream stage, imperfect visual perception would limit the effectiveness of structured abstraction; the detection errors would be accumulated from the perception stage to the abstraction stage. So visual perception is the basis of structured abstraction.

NN Learners with Perfect Perception About whether the results and claims about a neural-symbolic model can transfer to neural networks. We conducted a new experiment Table 3 to show the QA results of two representative NN learners (ViusalBERT [7] and ClipBERT [6]). We simulate perfect visual perceptions by using object bounding box ground-truth features.

The overall performances with GT are about 29% to 36% accuracy, and the results on the relatively hard question types (prediction/feasibility) are lower than easy ones (interaction/sequence) with about 3% to 8% decrease. Such results illustrate the NN models (without structured abstraction and reasoning) struggle to answer the reasoning questions, especially for harder tasks.

6 STAR-Humans

We add a STAR-Humans subset with 400 free-form human language questions for the four question types. Like CLEVR-Humans [3], the questions are primed with questions from and restricted to answers in STAR but do not need prior knowledge. Some questions are rewordings from STAR, others have altered semantics. We tested the program parser (pretrained on STAR) on the questions. Performances on four questions are 43.78%, 37.96%, 39.19%, 41.27%. The program parser potentially adapts word-difference or form-similar questions to executable programs, but can not handle the questions when the predicted programs are not executable.

Method Setting	Object Detection GT	Relationship Detection GT	Hypergraph Prediction GT	Language Parsing GT
oracle version (all GT)	✓	✓	✓	✓
w/o perfect hypergraphs (Obj GT, Rel GT, Graph Det)	✓	✓	✗	✓
w/o perfect visual perception (Obj GT, Rel Det)	✓	✗	✗	✓
w/o perfect visual perception (Obj Det, Rel Det)	✗	✗	✗	✓
w/o perfect language understanding (Graph GT)	✓	✓	✓	✗
w/o GT	✗	✗	✗	✗

Table 4: Variant Settings of Outcome-controlled Experiments in Diagnostic Model Evaluation

Interaction Question

Q: Which did person do with the broom?
A: Took.

```
Equal(
  Query_Objs(
    Unique(
      Filter_Actions_with_Obj(
        Actions( Filter_Situations_with_Obj
                  ( Situations( ), broom )
                ), broom
              )
            )
          ) , take
    )
```

Sequence Question

Q: Which object did the person take after they closed the book?
A: The dish.

```
Equal (
  Query_Objs(
    Filter_Actions_with_Verb(
      Filter_After_Actions(
        Actions( Situations( ) ), Unique(
          Filter_Actions_with_Verb(
            Actions(Filter_Situations_with_Obj
                      ( Situations( ), book )
                      ), close
                    )
                  ), take)
          ) , dish
    )
```

Prediction Question

Q: What will the person do next?
A: Close the door.

```
Belong_to(.
  Union ( close, door) , Query_Actions(
    Filter_After_Actions(
      Actions( Situations( ) )
        Query_Earliest_Action(
          Actions( Situations( ) )
          )
        )
      )
  )
```

Feasibility Question

Q: Which other object is possible to be put down by the person?
A: The clothes

```
Belong_to (
  Clothes, Except(
    Filter_Objs_by_Verb(
      Objs( Situations( ) ), put )
        Query_Objs(
          Query_Earliest_Action(
            Filter_Actions_with_Verb(
              Actions( Situations( ) )
            )
          ) , put
      )
  )
```

Figure 5: Functional program examples of four question types.

7 Dataset Examples

We show several examples of STAR benchmark in Figure 6 and Figure 7. Each situation video products multiple questions with generated answer, distractors and logic programs.

8 Data License

We use Charades videos under their Non-Commercial license and Action Genome frame annotations under its MIT license. Our dataset follows the Apache 2.0 license.

Data Type	Description	Instance Example
action	*a tuple contains action class, start and end time*	(c102, 15.0, 22.1)
verb	*the verb in the vocabulary*	take
object	*the object in the vocabulary*	laptop
relationship	*the relationship in the vocabulary*	over

Table 5: Basic data types for program execution.

Module Type	Function Name and Description	Inputs	Outputs
Input	**Situations** *Return all situations.*	\	*situations*
Element	**Actions** *Return all actions in the input situations.*	*situations*	*actions*
	Objs *Return all objects in the input situations.*	*situations*	*objects*
	Rels *Return all relationships in the input situations.*	*situations*	*relationships*
Filter	**Filter_Actions_with_Verb** *Select actions from input actions with the input verb.*	*(actions, verb)*	*actions*
	Filter_Actions_with_Obj *Select actions from input actions with the input object.*	*(actions, object)*	*actions*
	Filter_After_Actions *Select actions happened after input action.*	*(actions, action)*	*actions*
	Filter_Before_Actions *Select actions happened before input action.*	*(actions, action)*	*actions*
	Filter_Verbs_by_Obj *Select verbs that are compositional with the input object.*	*(verbs, object)*	*verbs*
	Filter_Objs_by_Verb *Select objects that are compositional with the input verb.*	*(objects, verb)*	*objects*
	Filter_Situations_with_Rel *Select situations that contain the input relationship.*	*(situations, relationship)*	*situations*
	Filter_Situations_with_Obj *Select situations that contain the input object.*	*(situations, object)*	*situations*
Query	**Query_Verbs** *Return verbs in the input actions.*	*actions*	*verbs*
	Query_Objs *Return objects in the input actions.*	*actions*	*objects*
	Query_Actions *Return action classes in the input actions.*	*actions*	*action_classes*
	Query_Earliest_Action *Return the action happened earliest in the input actions.*	*actions*	*action*
	Query_Latest_Action *Return the action happened latest in the input actions.*	*actions*	*action*
Logic	**Unique** *Return True if the set contains only one element, else return False.*	*set*	*bool*
	Equal *Return True if two input sets are equal, else return False.*	*(set1, set2)*	*bool*
	Belon_to *Return True if the set2 is included in the set1, else return False.*	*(set1, set2)*	*bool*
	Union *Return the union of the input sets.*	*(set1, set2)*	*set*
	Except *Return the except of the input sets.*	*(set1, set2)*	*set*

Table 6: The definitions and descriptions of program functions used in our program executor. Here *actions*, *verbs*, *objects* and *relationships* is the list of the corresponding basic data type. *situations* is the combination of the basic data type. And *set* represents the set of basic verbs, objects, relationships or actions.

Interaction_T1_2200
Q: What object was thrown by the person?
a. The clothes. b. The bag.
c. The blanket. d. The pillow.

Sequence_T4_1391
Q: What happened before the person put down the bag?
a. Closed the door. b. Put down the broom.
c. Threw the clothes. d. Tidied up the blanket.

Prediction_T1_843
Q: What will the person do next?
a. Sit on the sofa. b. Take the phone/camera.
c. Hold the laptop. d. Take the box.

Feasibility_T6_467
Q: What is the person able to do after walking through the doorway?
a. Close the box. b. Throw the bag.
c. Put down the dish. d. Open the door

Interaction_T2_3450
Q: What did the person do with the dish?
a. Put down. b. Washed.
c. Took. d.Lied on.

Sequence_T2_2544
Q: Which object did the person open before they took the cup?
a. The blanket. b. The refrigerator.
c. The box. d. The cabinet

Prediction_T1_1317
Q: What will the person do next?
a. Put down the box. b. Hold the laptop.
c. Put down the broom. d. Put down the sandwich.

Feasibility_T2_602
Q: What else is the person able to do with the cup?
a. Sit on the cup. b. Wash the cup.
c. Put down the cup. d. Take the cup

Interaction_T2_1807
Q: What did the person do with the clothes?
a. Put down. b. Took.
c. Washed. d. Tidied up.

Sequence_T6_1180
Q: What did the person do to the blanket before taking the clothes?
a. Put down. b. Threw.
c. Sat on. d. Took.

Prediction_T1_703
Q: What will the person do next?
a. Lie on the bed. b. Take the box.
c. Close the door. d. Put down the cup.

Feasibility_T2_356
Q: What else is the person able to do with the clothes?
a. Put down the clothes. b. Tidy up the clothes.
c. Take the clothes. d. Sit on the clothes

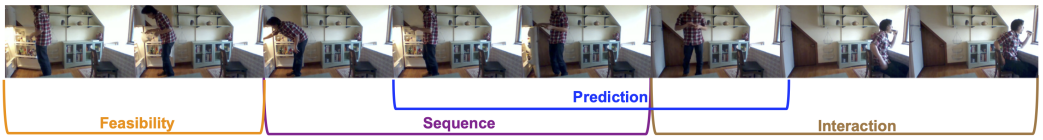
Figure 6: Examples from STAR. Green choices are ground-truth answers, and red choices are incorrect options. Best viewed in color.

Interaction_T1_3751
Q: What object was thrown by the person?
a. The broom. b. The towel.
c. The blanket. d. The shoe.

Sequence_T1_3354
Q: Which object did the person take after they closed the refrigerator?
a. The food. b. The phone.
c. The towel. d. The bottle.

Prediction_T2_505
Q: What will the person do next with the refrigerator?
a. Wash. b. Tidy up.
c. Open. d. Close.

Feasibility_T2_597
Q: What else is the person able to do with the refrigerator?
a. Eat the refrigerator. b. Put down the refrigerator.
c. Open the refrigerator. d. Close the refrigerator.

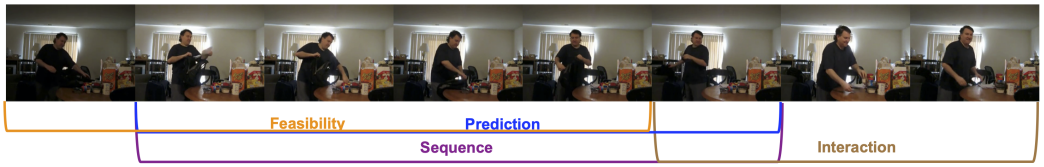


Interaction_T1_703
Q: Which object was sat at by the person?
a. The table. b. The sofa.
c. The floor. d. The bed.

Sequence_T1_708
Q: Which object did the person close after they held the food?
a. The book. b. The laptop.
c. The box. d. The refrigerator.

Prediction_T2_205
Q: Which object would the person eat next after they close the refrigerator?
a. The sandwich. b. The paper.
c. The bag. d. The medicine.

Feasibility_T5_295
Q: Which object is the person able to take after opening the refrigerator?
a. The laptop. b. The broom.
c. The clothes. d. The food.



Interaction_T2_6123
Q: What did the person do with the table?
a. Washed. b. Sat on.
c. Put down. d. Tidied up.

Sequence_T6_3936
Q: What did the person do to the notebook before throwing the bag?
a. Sat on. b. Took.
c. Put down. d. Closed.

Prediction_T4_1467
Q: Which object would the person tidy up next after they put down the notebook?
a. The blanket. b. The broom.
c. The table. d. The towel.

Feasibility_T2_980
Q: What else is the person able to do with the bag?
a. Close the bag. b. Throw the bag.
c. Put down the bag. d. Open the bag.

Figure 7: Examples from STAR. Green choices are ground-truth answers, and red choices are incorrect options. Best viewed in color.

9 Datasheets for Dataset

In this section, we provide the datasheets for the STAR benchmark.

9.1 Motivation

The questions in this section are primarily intended to encourage dataset creators to clearly articulate their reasons for creating the dataset and to promote transparency about funding interests.

- **For what purpose was the dataset created?** Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.
  We would like to introduce a new benchmark for situated reasoning, evaluating multiple situated reasoning ability from real-world videos through logic-grounded question answering.
- **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**
  STAR team (Bo Wu) in MIT-IBM Watson AI Lab.
- **Who funded the creation of the dataset?** If there is an associated grant, please provide the name of the grantor and the grant name and number.

9.2 Composition

Dataset creators should read through the questions in this section prior to any data collection and then provide answers once collection is complete. Most of these questions are intended to provide dataset consumers with the information they need to make informed decisions about using the dataset for specific tasks. The answers to some of these questions reveal information about compliance with the EU's General Data Protection Regulation (GDPR) or comparable regulations in other jurisdictions.

- **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.
  Each sample of STAR associates with a question, answer, three candidate options, situation hypergraph, logic program, annotations (entity bounding boxes, classes), video clip IDs, and video keyframe IDs .
- **How many instances are there in total (of each type, if appropriate)?**
  STAR consists of 60.9K situated reasoning questions, 243.6K candidate choices, 140.7K structured situation hypergraphs, and 23.2K trimmed situation video clips. Situations in our benchmark are sourced from Charades videos in daily-life scenes.
- **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).
  The videos and question answer pairs in STAR are generated by machines. In theory, we can generate more instances with more source videos. We release all the instances we have synthesized.
- **What data does each instance consist of?** "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.
  Each sample of STAR associates with a generated question, answer, three candidate options, situation hypergraph, logic program, annotations (entity bounding boxes, classes), and reference video clip IDs.
- **Is there a label or target associated with each instance?** If so, please provide a description.
  Yes. We provide question answering labels for objects' attributes, locations and physical properties in the videos. For each question answer pairs, we also provide their underlying functional programs.

- **Is any information missing from individual instances?** If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

  No.

- **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?** If so, please describe how these relationships are made explicit.

  No.

- **Are there recommended data splits (e.g., training, development/validation, testing)?** If so, please provide a description of these splits, explaining the rationale behind them.

  We provide data split file in our homepage (STAR) [1]

- **Are there any errors, sources of noise, or redundancies in the dataset?** If so, please provide a description.

  No.

- **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?** If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

  This dataset provides video IDs from the Charades dataset under their Non-Commercial license and keyframe from the Action Genome under its MIT license.

- **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)?** If so, please provide a description.

  No.

- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** If so, please describe why.

  No.

- **Does the dataset relate to people?** If not, you may skip the remaining questions in this section.

- **Does the dataset identify any subpopulations (e.g., by age, gender)?** If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

  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?** If so, please describe how.

  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)?** If so, please provide a description.

  No.

9.3 Collection Process

As with the previous section, dataset creators should read through these questions prior to any data collection to flag potential issues and then provide answers once collection is complete. In addition to the goals of the prior section, the answers to questions here may provide information that allow

[1] STAR Benchmark: http://star.csail.mit.edu or https://bobbywu.com/STAR

others to reconstruct the dataset without access to it.

- **How was the data associated with each instance acquired?** Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

  We generated instances in STAR that with corresponding video ,question ,answer and situation graphs from Charades and Action Genome. And we use Charades under their Non-Commercial license and Action Genome under its MIT license.

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

  We design a QA generation diagram to generate data each instance from Charades and Action Genome. The code is released on our homepage.

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

  No.

- **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)?**

  We submit labeling task on Amazon Mechanical Turk. We spend 500 dollars in total.

- **Over what timeframe was the data collected?** Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

  Our dataset is generated from Charades and Action Genome. We generate the dataset from January 2021 to July 2021.

- **Were any ethical review processes conducted (e.g., by an institutional review board)?** If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

  Yes. We provide a Broader Impact section in the supplementary.

- **Does the dataset relate to people?** If not, you may skip the remainder of the questions in this section.

  Yes.

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

  No. Our video data is from Charades under their Non-Commercial license. Charades Homepage: https://prior.allenai.org/projects/charades.

- **Were the individuals in question notified about the data collection?** If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

  Not applicable. Our data is from Charades under their Non-Commercial license and Action Genome under its MIT license.

- **Did the individuals in question consent to the collection and use of their data?** If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

  Not applicable. Our data is generated from Charades under their Non-Commercial license and Action Genome under its MIT license

- **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?** If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

  Not applicable.

- **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?** If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

  No.

- **Any other comments?**

  STAR is a generated dataset from Charades under their Non-Commercial license and Action Genome under MIT license. The collection process of raw data should refer to the sources.

9.4 Preprocessing/cleaning/labeling

Dataset creators should read through these questions prior to any preprocessing, cleaning, or labeling and then provide answers once these tasks are complete. The questions in this section are intended to provide dataset consumers with the information they need to determine whether the "raw" data has been processed in ways that are compatible with their chosen tasks. For example, text that has been converted into a "bag-of-words" is not suitable for tasks involving word order.

- **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)?** If so, please provide a description. If not, you may skip the remainder of the questions in this section.
- **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?** If so, please provide a link or other access point to the "raw" data.
- **Is the software used to preprocess/clean/label the instances available?** If so, please provide a link or other access point.

9.5 Uses

These questions are intended to encourage dataset creators to reflect on the tasks for which the dataset should and should not be used. By explicitly highlighting these tasks, dataset creators can help dataset consumers to make informed decisions, thereby avoiding potential risks or harms.

- **Has the dataset been used for any tasks already?** If so, please provide a description.

  Yes. It has been used for Situated Reasoning.

- **Is there a repository that links to any or all papers or systems that use the dataset?** If so, please provide a link or other access point.

  Yes. We include the paper link and the repository link in our homepage.

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

  It includes Video Question Answering, Video Commonsense Reasoning.

- **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

  No.

- **Are there tasks for which the dataset should not be used?** If so, please provide a description.

  Not applicable.

9.6 Distribution

Dataset creators should provide answers to these questions prior to distributing the dataset either internally within the entity on behalf of which the dataset was created or externally to third parties.

- **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?** If so, please provide a description.

  STAR is a academic dataset for public non-commercial use.

- **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?** Does the dataset have a digital object identifier (DOI)?

  It is released on a website and Github.

- **When will the dataset be distributed?**

  After the paper was published.

- **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

  Yes, Apache 2.0.

- **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

  No.

- **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

  No.

9.7 Maintenance

As with the previous section, dataset creators should provide answers to these questions prior to distributing the dataset. These questions are intended to encourage dataset creators to plan for dataset maintenance and communicate this plan with dataset consumers.

- **Who is supporting/hosting/maintaining the dataset?**

  Bo Wu and Shoubin Yu.

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

  bobbywu.cs@gmail.com

- **Is there an erratum?** If so, please provide a link or other access point.

  No.

- **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?** If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

  If the dataset is updated we will update it the on our homepage.

- **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)?** If so, please describe these limits and explain how they will be enforced.

  No.

- **Will older versions of the dataset continue to be supported/hosted/maintained?** If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

  Yes.

- **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

  Yes. We provide our data generation coded in our homepage.

References

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[2] J. Ji, R. Krishna, L. Fei-Fei, and J. C. Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *CVPR*, 2020.

[3] J. Johnson, B. Hariharan, L. Van Der Maaten, J. Hoffman, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick. Inferring and executing programs for visual reasoning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2989–2998, 2017.

[4] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[5] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 2017.

[6] J. Lei, L. Li, L. Zhou, Z. Gan, T. L. Berg, M. Bansal, and J. Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7331–7341, 2021.

[7] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[8] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[9] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[10] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Springer, Cham*, 2016.

[11] K. Tang, Y. Niu, J. Huang, J. Shi, and H. Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3716–3725, 2020.

[12] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. *arXiv preprint arXiv:1611.05431*, 2016.