ESPRIT: Explaining Solutions to Physical ReasonIng Tasks

Rui Zhang
Yale University

LILY Lab

Motivation

- Humans learn to understand and reason about physical laws just by living in this world and doing everyday things.
- AI models, on the other hand, lack this ability and so are unable to generalize to new scenarios by reasoning about abstract physical concepts like gravity, mass, inertia, friction, and collisions
- We propose ESPRIT, a framework for commonsense reasoning about qualitative physics in natural language that generates interpretable descriptions of physical events.

PHYRE Benchmark Dataset

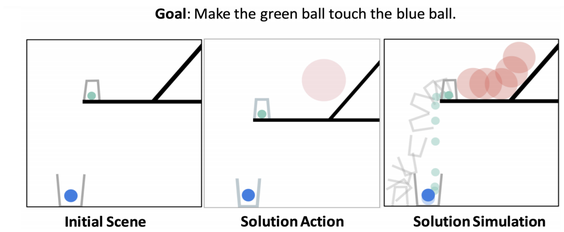
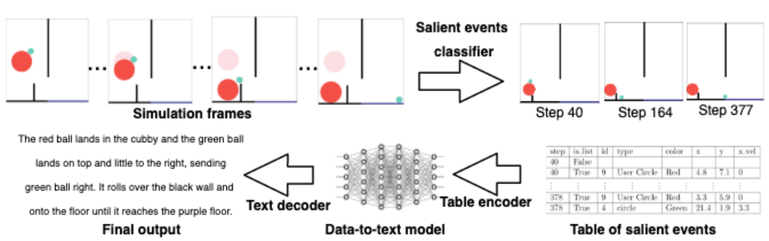


Figure 1: An example of PHYRE benchmark (Bakhtin et al., 2019) consisting of a goal and an initial scene, its solution action, and solution simulation. Each object color corresponds to an object type. Red: the user-added dynamic object; Green and Blue: dynamic goal object; Purple: the static goal object; Gray: the dynamic scene object; Black: the static scene object.

ESPRIT Framework



ESPRIT Dataset

Templates	25
Tasks	2441
Objects / Task	13.6
Frames / Task	657.9
Collisions / Task	54.2
Annotated Tasks (train/dev/test)	625/84/76
Collisions / Annotated Task	24.5
Important Collisions / Annotated Task	3.9
Tokens / Initial State Description	38
Tokens / Simulation Description	44
Vocabulary Size	867

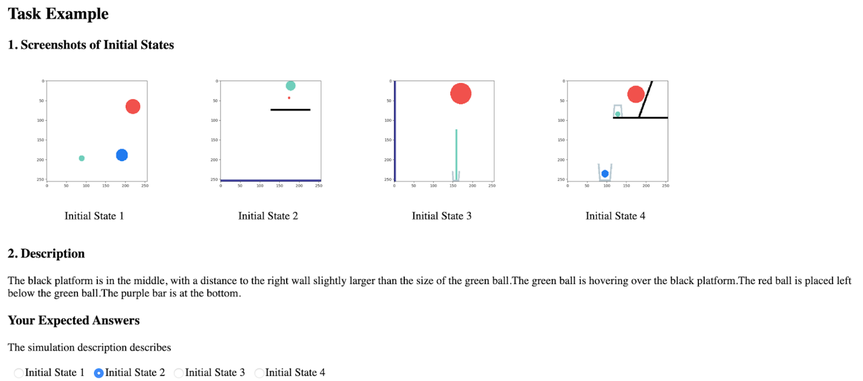
Table 1: Statistics for the ESPRIT Dataset.

Data-to-Text Generation

Input records	... green|green_circle_0|OBJ_COLOR| INITIAL_STATE circle|green_circle_0|OBJ_TYPE| INITIAL_STATE dynamic|green_circle_0|OBJ_STATE| INITIAL_STATE 76|green_circle_0|X| INITIAL_STATE 162|green_circle_0|Y| INITIAL_STATE...
Gold annotation	The red and green balls fall. The red ball lands on the ground and the green ball lands on the red ball and rolls to the right over the black vertical bar.
Generation (AVG)	The red ball lands in the cubby and the green ball lands on top and a little to the right, sending the green ball right. It rolls over the short black wall of the cage and onto the floor, where it keeps rolling right towards the purple goal...
Generation (BiLSTM)	The red ball falls and knocks the green ball off of its curved black platform and to the left. It rolls leftwards and continues falling until it lands on the purple floor...

Table 6: Example input records, gold annotation and generated simulation description from the AVG and BiLSTM models. This example is taken from 00014:394. We show only a short segment of the actual input records.

Human Evaluation of Validity and Coverage



Experimental Results

	Initial state	Simulation
Random classifier	25.0	25.0
GPT (Radford et al., 2018)	14.8	44.4
AVG (Puduppully et al., 2019b)	85.2	74.1
BiLSTM (Puduppully et al., 2019b)	81.5	51.9
Human Annotation	66.7	63.0

Table 4: Human evaluation for *validity* accuracy of initial state and simulation descriptions on test set.

	Gravity	Friction	Collision
GPT (Radford et al., 2018)	3.9	0.0	6.6
AVG (Puduppully et al., 2019b)	100.0	96.1	86.8
BiLSTM (Puduppully et al., 2019b)	100.0	93.4	84.2
Human Annotation	94.7	57.9	51.3

Table 5: Human evaluation for *coverage* accuracy of physical concepts in simulation descriptions on test set.