



# Knowledge-Grounded Self-Rationalization via Extractive and Natural Language Explanations

Bodhisattwa Prasad Majumder, Oana-Maria Camburu, Thomas Lukasiewicz, Julian McAuley  
UC San Diego, University of Oxford, TU Wien Informatics

**ICML 2022**



**ICML**  
International Conference  
On Machine Learning

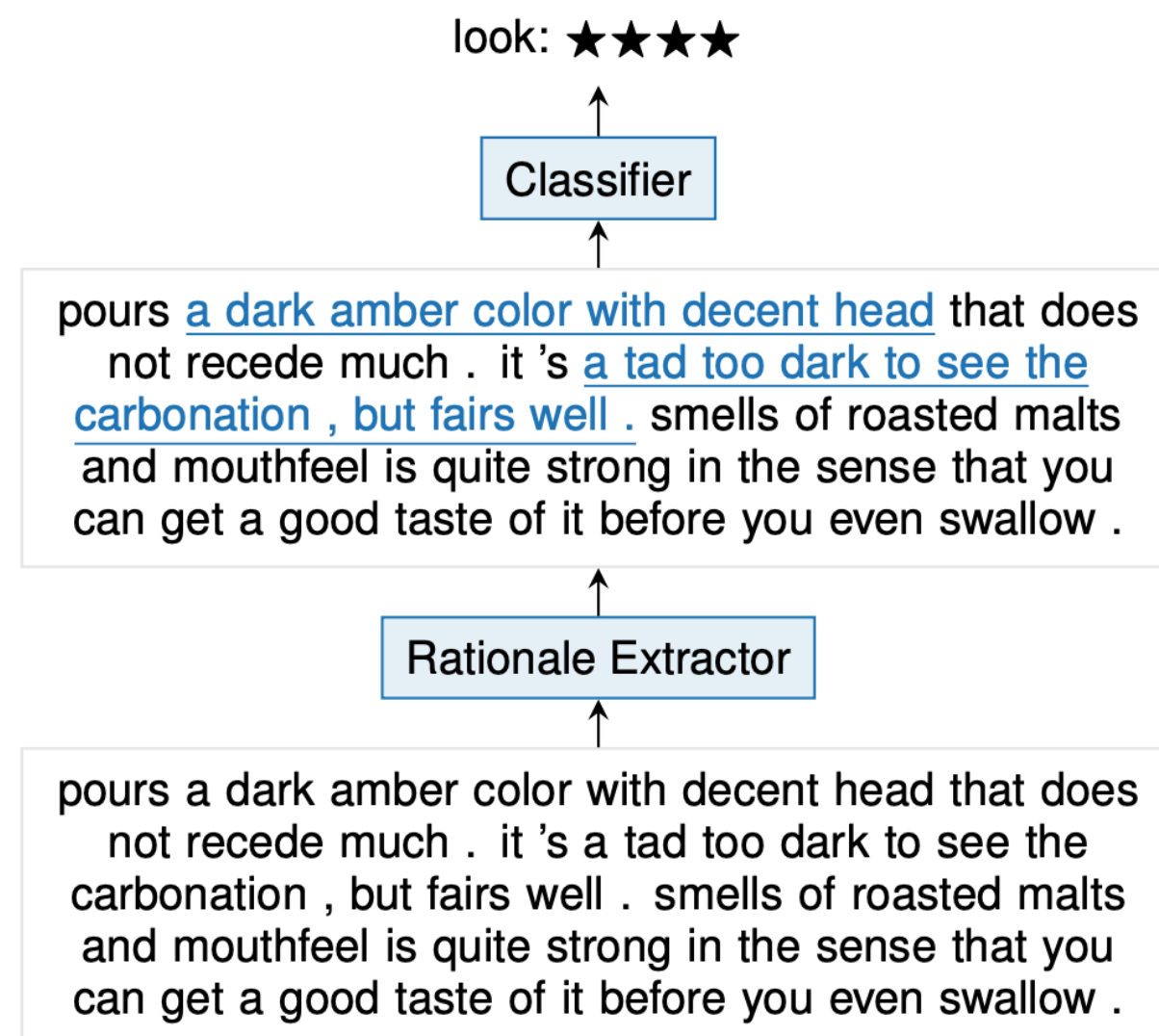




# User Experience with AI Explanations



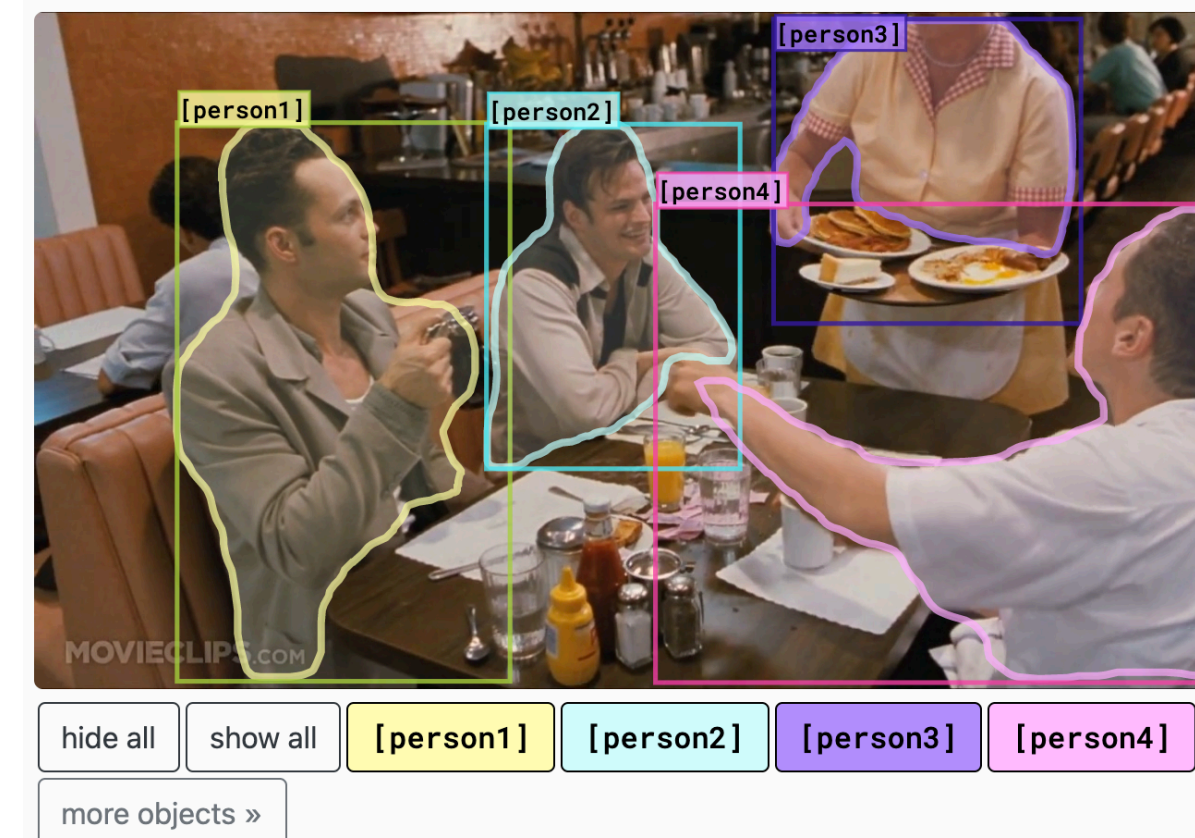
Selvaraju et al., 2019



Bastings et al., 2020

2210	1	617	653	23903	497	21596	26688	4519	4190	19999	47	1	20	11	1	3066	276	466	2	2116	367	1435	3008	1	1	3096	89	175	13730	226	796	107	12802	867	10174	8056	1770	1313	11763	7995	2687	3	154	5065	551	14140	1			
12896	1	618	653	23903	498	13708	15862	2315	2195	14741	381	164	48	7	1	3164	398	466	1	1763	151	217	1644	1	1	3269	266	2183	759	824	42	3630	265	10476	1471	231	3672	2445	1263	2668	3	154	5065	551	14140	1				
4037	1	619	653	23903	499	13708	15862	2315	2195	14741	381	164	48	7	1	3164	398	466	1	1763	151	217	1644	1	1	3269	266	2183	759	824	42	3630	265	10476	1471	231	3672	2445	1263	2668	3	154	5065	551	14140	1				
7959	1	620	653	23903	500	13708	15862	2315	2195	14741	381	164	48	7	1	3164	398	466	1	1763	151	217	1644	1	1	3269	266	2183	759	824	42	3630	265	10476	1471	231	3672	2445	1263	2668	3	154	5065	551	14140	1				
1725	2	2578	1	22120	40	4261	12339	367	4062	3712	240	43	7	37	6	1	8105	38	213	1	1893	116	118	2634	1	1	8335	151	2695	1211	51	749	42	3120	805	2395	119	2019	155	1180	2156	211	3068	1	1	689	7740	1130	1	
11395	1	628	1228	22645	88	5986	14844	334	3071	2376	134	172	27	2	1	5452	128	294	1	3091	75	37	4056	18	18	5418	381	2292	1448	48	721	30	3343	63	2681	90	10971	126	1482	18437	201	4402	3	913	5648	1033	1			
3095	14	1929	1	21750	382	3056	10241	351	816	1655	258	102	15	5	1	3437	258	1	1	3657	19	72	4078	1	1	3545	85	2292	1443	113	831	131	2831	283	2645	33	10274	137	1729	1641	187	3624	1	1	568	809	930	1		
2049	1	631	1228	22645	89	5986	14844	334	3071	2376	134	172	27	2	1	5452	128	294	1	3091	75	37	4056	18	18	5418	381	2292	1448	48	721	30	3343	63	2681	90	10971	126	1482	18437	201	4402	3	913	5648	1033	1			
2144	4	466	2	18655	80	5673	801	380	350	1658	255	1249	9	24	18	21	3604	1	2192	25	1	7843	62	23	2245	38	2676	322	566	137	368	469	2137	35	17533	140	3206	1768	168	8558	1	1	408	6290	367	1				
13026	1	639	1	30977	119	1151	11226	139	1016	1215	267	1292	7	24	28	26	3564	292	262	5	1395	12	18	6742	18	37	2233	1	2734	717	433	410	112	5751	300	2571	65	10369	302	1250	36661	56	10578	8	2	71	7511	69	1	
1534	2	169	1	16215	790	3605	3060	118	1161	1132	172	1	1	1	1	1	1791	1	1	1	6391	74	48	2294	1	1	4007	629	675	89	109	10	5008	1237	1211	60	1237	77	1432	1137	9265	1	1	410	5587	75	1			
2338	1	641	1228	22645	90	5986	14844	334	3071	2376	134	172	27	2	1	5452	128	294	1	3091	75	37	4056	18	18	5418	381	2292	1448	48	721	30	3343	63	2681	90	10971	126	1482	18437	201	4402	3	913	5648	1033	1			
1127	1	64	1	21935	401	1390	2684	112	526	605	339	295	9	22	8	1	3412	9	344	1	1442	23	11	6391	74	48	2294	1	1	4007	629	675	89	10	5008	1237	1211	60	1237	77	1432	1137	9265	1	1	410	5587	75	1	
1506	1	61	1	18607	71	1500	2320	55	4017	808	554	222	8	17	5	1	1244	8	252	1	1178	20	1	7321	63	282	867	6	5902	35	787	347	40	1040	3900	1433	4125	40	7421	63	1237	5285	12	1441	1	1	258	4708	100	1
4372	1	62	1	14884	350	1451	101	4737	381	354	204	77	13	1	1	1	1	1	1	891	12	1	6391	74	48	2294	1	1	4007	629	675	89	10	5008	1237	1211	60	1237	77	1432	1137	9265	1	1	410	5587	75	1		
793	1	63	1	16329	403	1227	126	354	180	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		
649	9	60	1	18796	305	1161	902	309	207	302	291	113	7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1			
647	4	64	1	18925	226	640	1551	338	4754	260	387	8	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1			
428	7	6	1	18444	263	220	403	344	441	238	259	9	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1			
238	1	65	1	15336	435	1343	156	578	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		
788	7	6	1	22461	137	241	367	199	4142	288	173	102	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1			
309	15	6	68	16738	180	67	242	2729	249	203	138	9	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1			
239	12	1	68	15576	96	11	97	134	178	205	103	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1			
793	1	63	1	16329	403	1227	126	354	180	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1			
207	15	6	68	16738	180	67	242	2729	249	203	138	9	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1			
130	15	6	68	16738	180	67	242	2729	249	203	138	9	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1			
85	14	1	68	16738	180	67	242	2729	249	203	138	9	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1			
18	1	64	1	15918	35	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		
208	17	2	65	154	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
237	17	2	68	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
237	17	2	68	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
237	17	2	68	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
237	17	2	68	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
237	17	2	68	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
237	17	2	68	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
237	17	2	68	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
237	17																																																	

<https://jalammar.github.io/explainable-ai/>



Why is [person4] pointing at [person1]?

- a) He is telling [person3] that [person1] ordered the pancakes.
- b) He just told a joke.
- c) He is feeling accusatory towards [person1].
- d) He is giving [person1] directions.

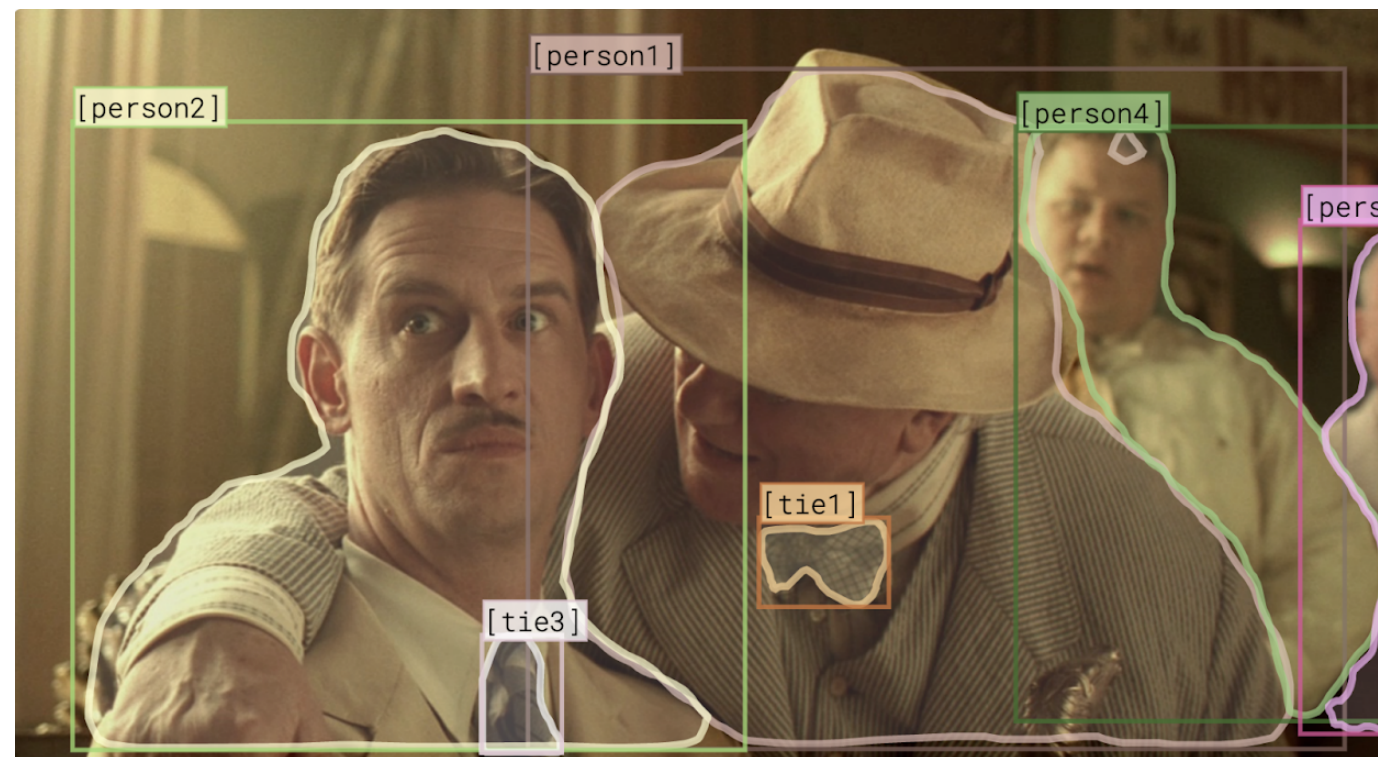
*Rationale: I think so because...*

- [person1]** has the pancakes in front of him.
- [person4]** is taking everyone's order and asked for clarification.
- [person3]** is looking at the pancakes both she and **[person2]** are smiling slightly.
- [person3]** is delivering food to the table, and she might not know whose order is whose.

Zellers et al., 2019



# Rich Representation of Explanations

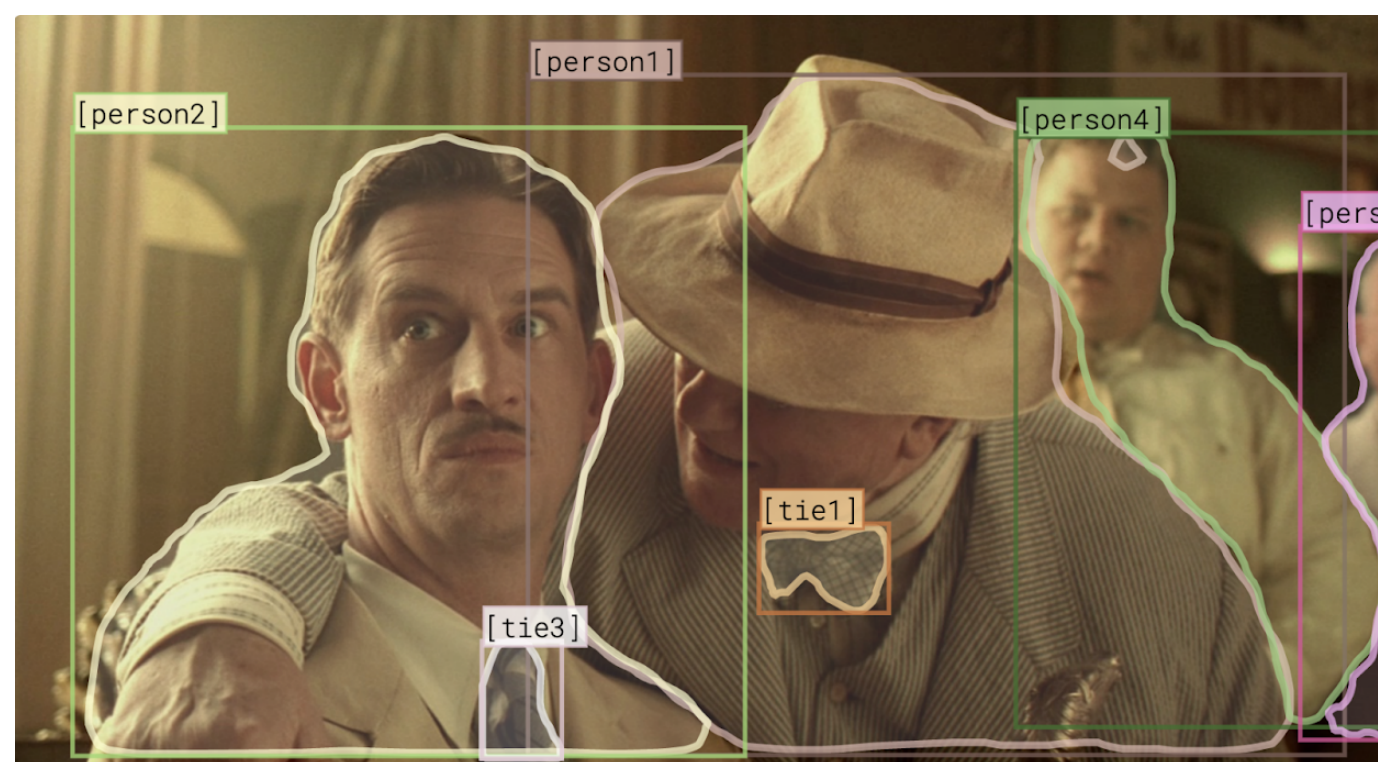


**Q:** how does  
[person2] feel about  
what [person1] is  
telling him?

**A:** He's concerned  
and a little upset



*extractive*



**Q:** how does  
[person2] feel about  
what [person1] is  
telling him?

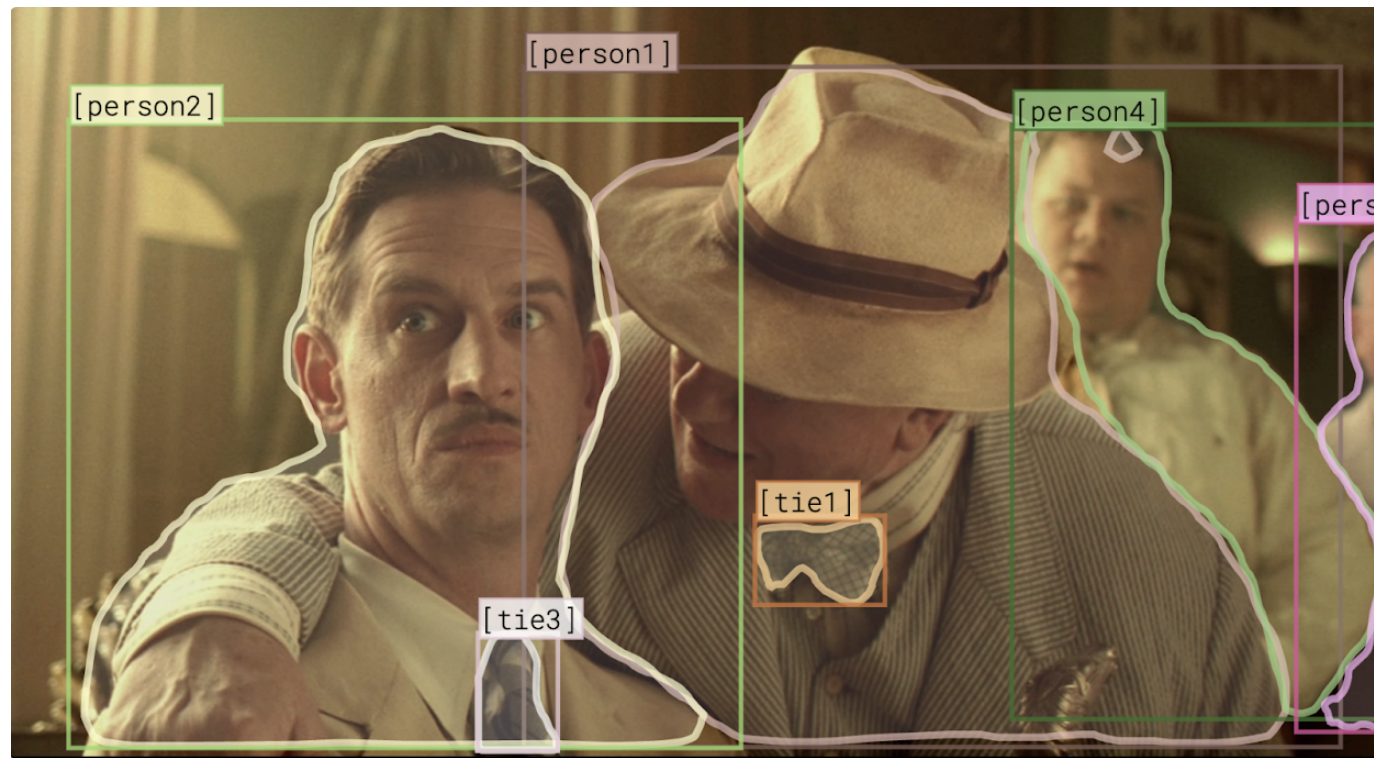
**A:** He's concerned  
and a little upset

He is in shock thinking  
something bad is about  
to happen.

*abstractive*



# Natural Language Explanations (NLEs)



**Q:** how does  
[person2] feel about  
what [person1] is  
telling him?

**A:** He's concerned  
and a little upset

He is in shock thinking  
something bad is about  
to happen.

*abstractive*

- NLE should be **plausible** and consistent to the input
- NLE should be **accurate** and **faithful** to explain the prediction
- NLE should be grounded into **world knowledge**



# Predictive Task

A neural predictive model is employed to solve task.

For example: **Natural Language Inference (NLI)**

**premise**

Two men are competing in a  
bicycle race

**hypothesis**

People are riding bikes

**label**  
*entailment*

**Instance from SNLI dataset**



# Rationales

A rationale (or extractive rationale) is a sufficient and minimal part of the **input** that is a **significant indicator** of a model's prediction.

**premise**

Two men are competing in a  
bicycle race

**hypothesis**

People are riding bikes

**label**  
*entailment*

Realized via smallest lexical units e.g., tokens for language or super-pixels for images



# Natural Language Explanations (NLEs)

An NLE is a **textual abstraction** of the model explanation. This is grounded in background knowledge that the model believes.

**premise**

Two men are competing in a  
bicycle race

**hypothesis**

People are riding bikes

**label**  
***entailment***

Competing in a  
bicycle race  
requires men  
riding bikes



# Background Knowledge

A model **believes** in a set of background knowledge given the input. This knowledge is pivotal to construct the NLEs.

**premise**

Two men are competing in a  
bicycle race

**hypothesis**

People are riding bikes

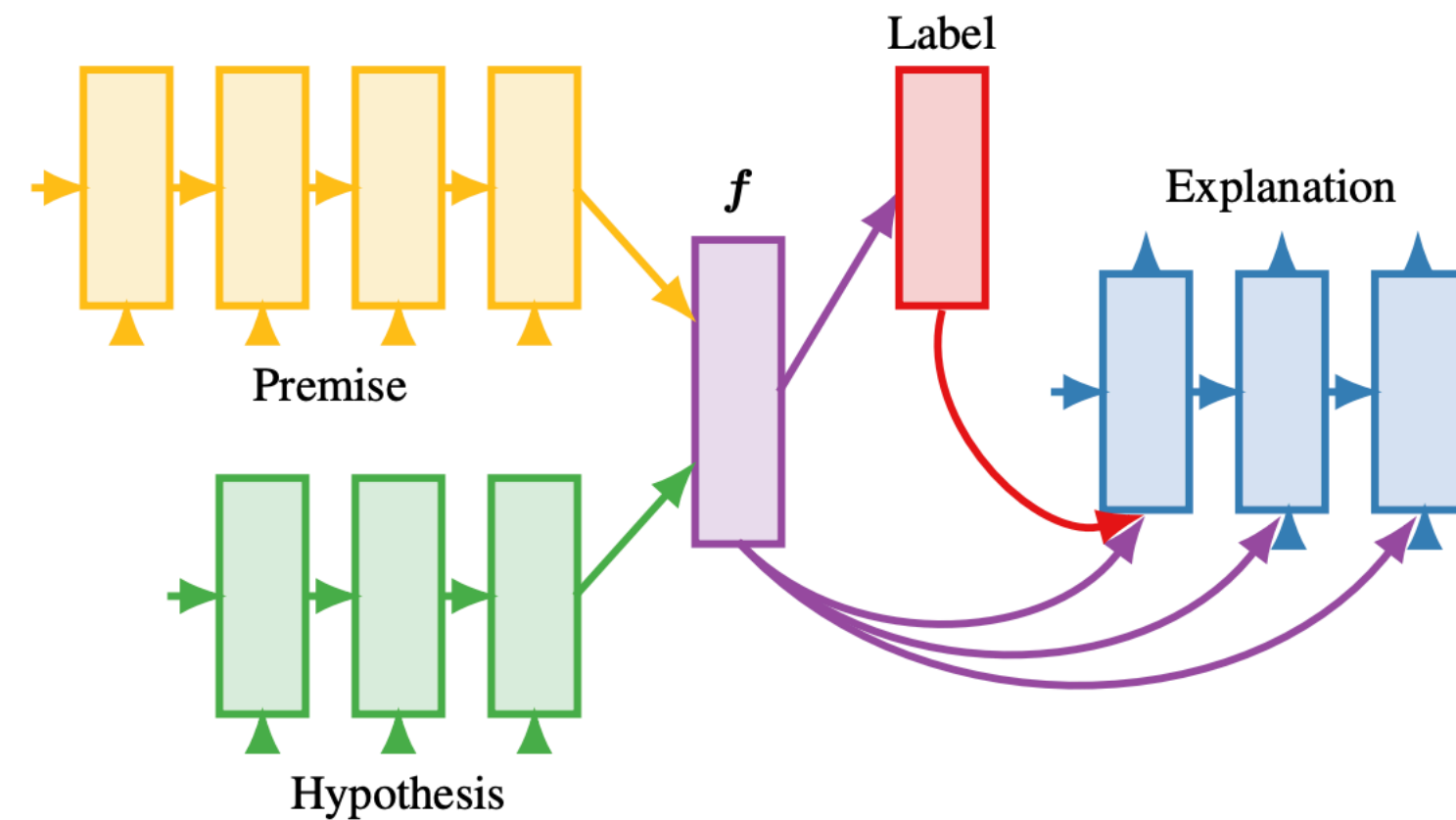
- bicycle race requires bikes
- race requires riding bikes
- bicycle race needs helmet
- men are people

**label**  
***entailment***

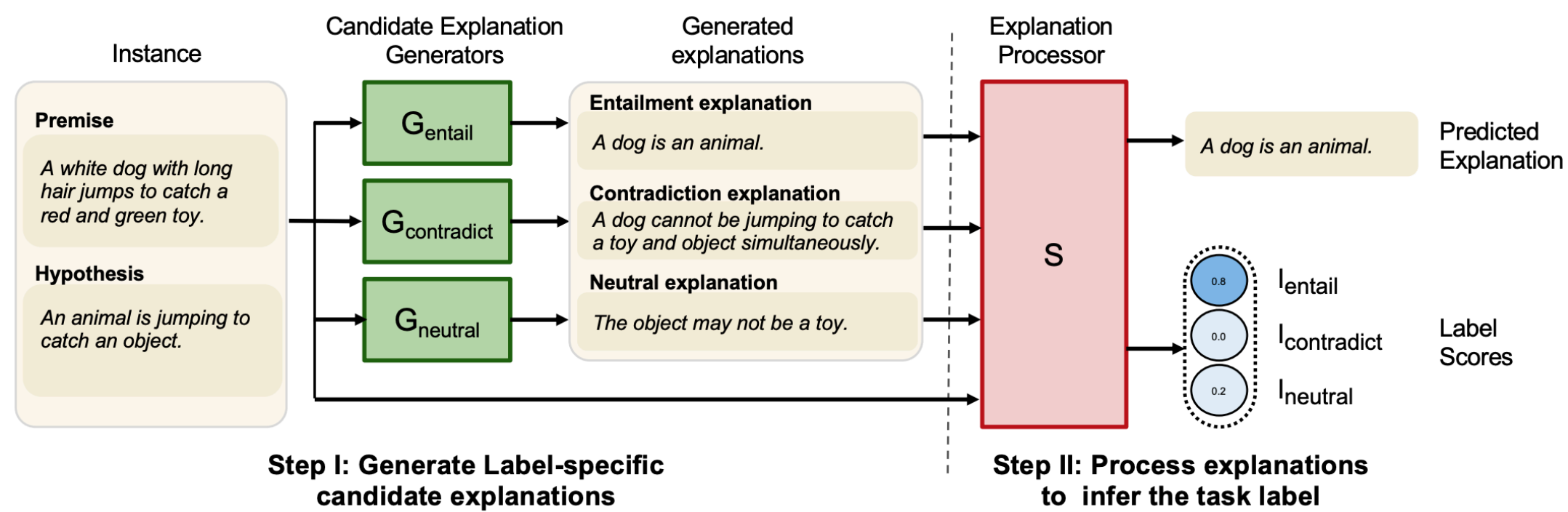
Competing in a  
bicycle race  
requires men  
riding bikes



# Previous Works



*predict-then-explain* (Camburu et al., 2018)



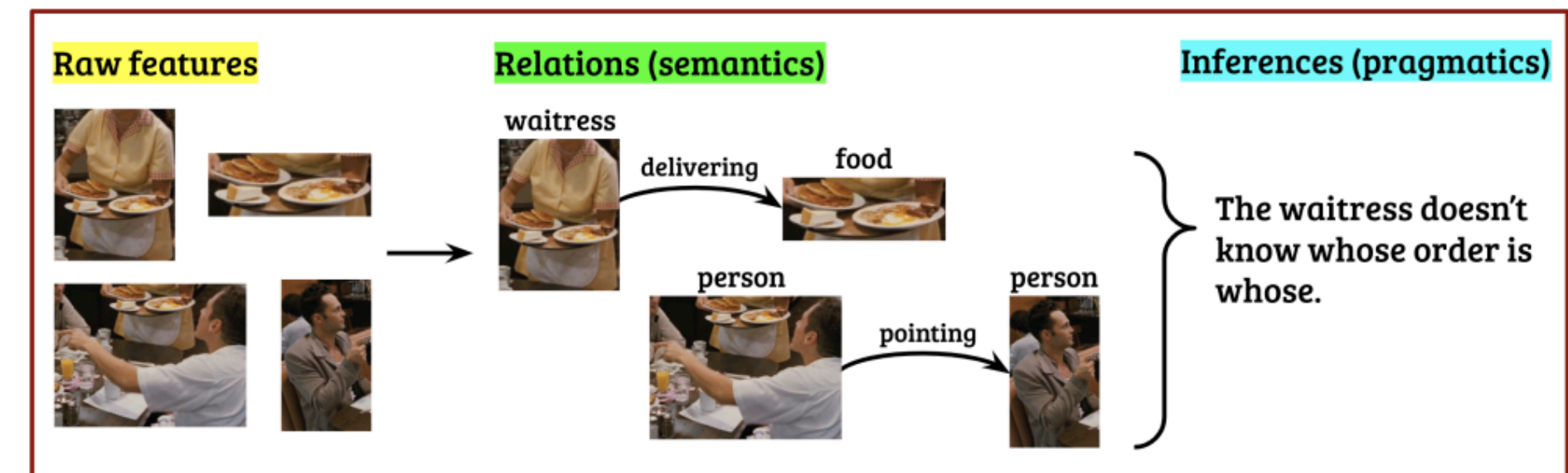
*generate label-specific explanations, then choose the correct one*  
(Kumar et al., 2018)



**Question:** Why is person on the right pointing to the person on the left?

**Answer:** He is telling the waitress that the person on the left ordered the pancakes.

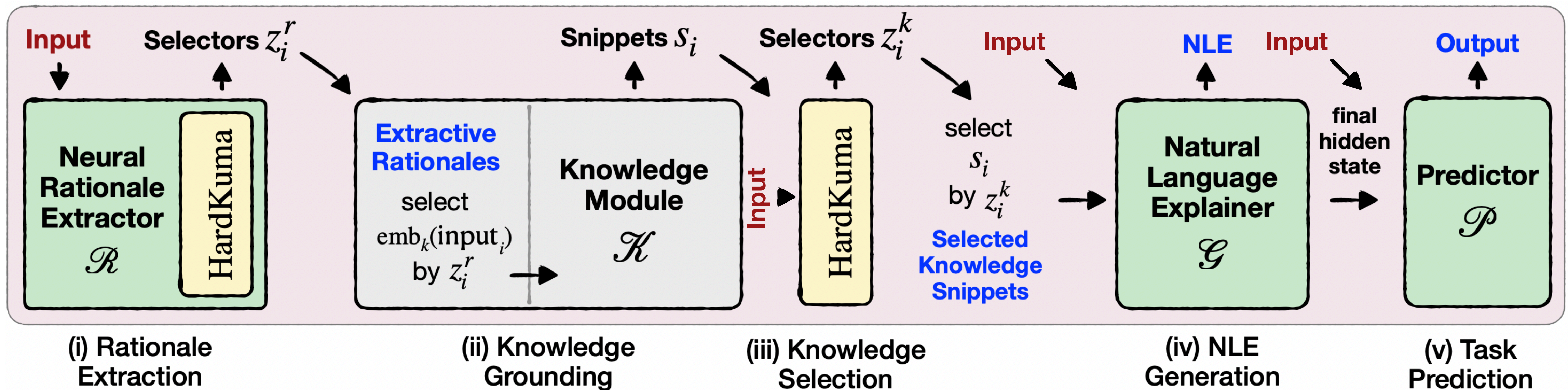
**Natural language rationale:** The answer is true because she is delivering food to the table and she doesn't know whose order is whose.



*stacked steps of feature extraction, selection, commonsense inference*  
(Marasovic et al., 2018)



# Rationale + Knowledge + NLE = RExC



Rationales are responsible for relevant knowledge retrieval

Knowledge (latent) selection acts as a **soft bottleneck**

RExC is a **self-rationalizing** model that produces NLE and task output



# Natural Language and Visual-Language Tasks

Natural Language Tasks

Natural Language Inference

premise

Two men are competing in a bicycle race

hypothesis

People are riding bikes

label  
*entailment*

Commonsense Validation

A: Coffee stimulates people

B: Coffee depresses people

label  
*B is invalid*

Commonsense QA

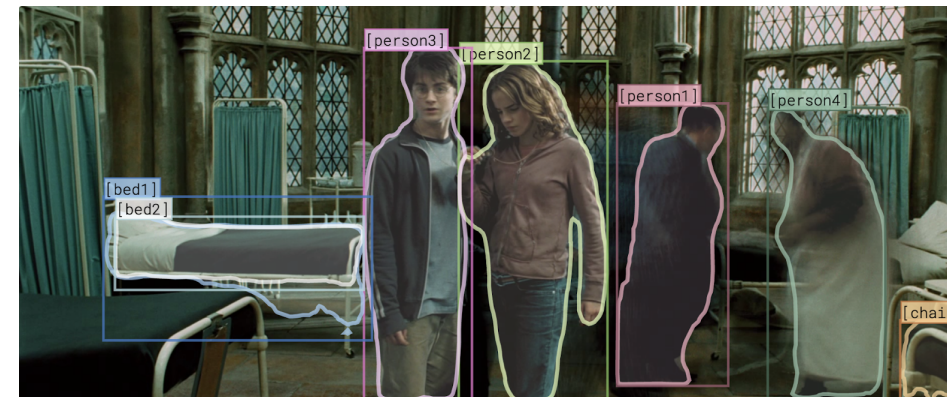
Q: Where does a wild bird usually live?

A: a) cage, b) sky, c) countryside, d) desert, e) windowsill

label  
*sky*

Vision Language Tasks

Visual Entailment



Hypothesis:  
Some tennis players pose

label  
*entailment*

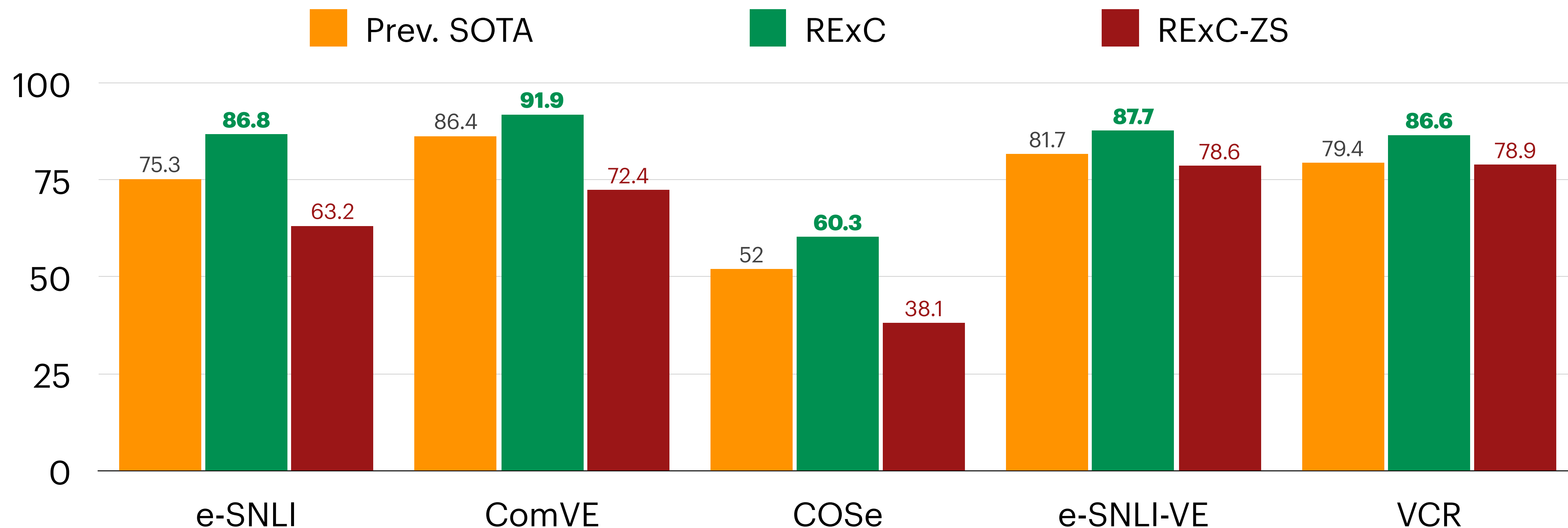
Visual Commonsense Reasoning



Q: What is the place?

label  
*They are in a hospital room*

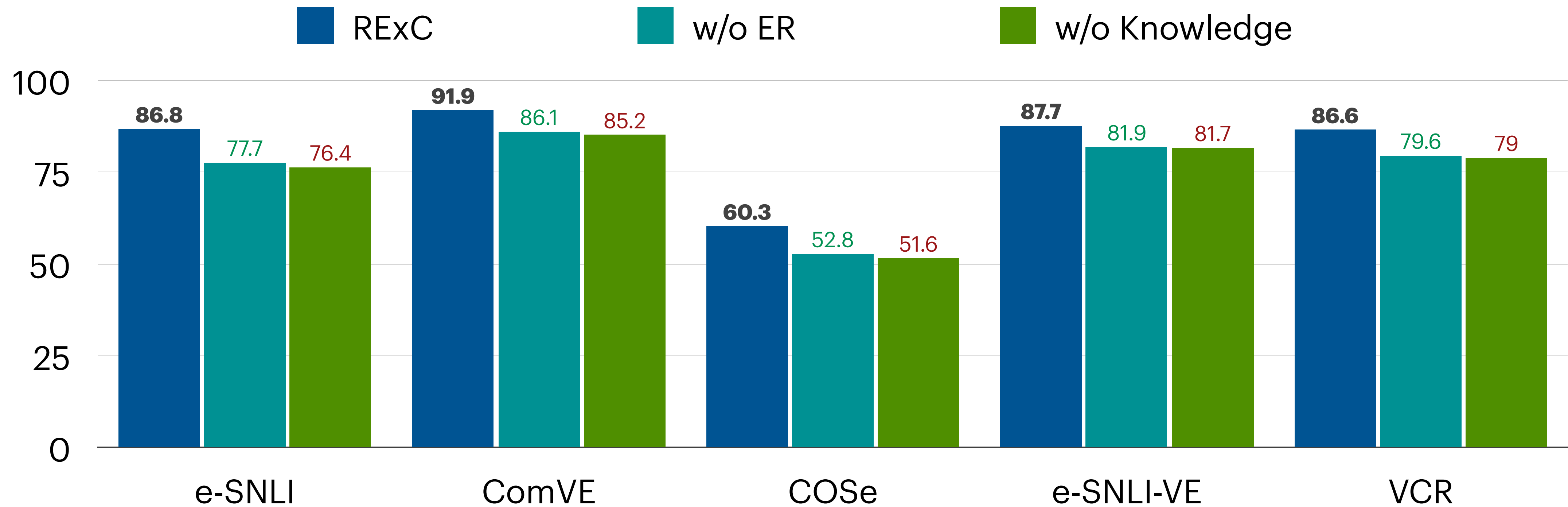
# Plausibility via BertScore



RExC **outperforms** all previous SOTA for NLE quality/plausibility

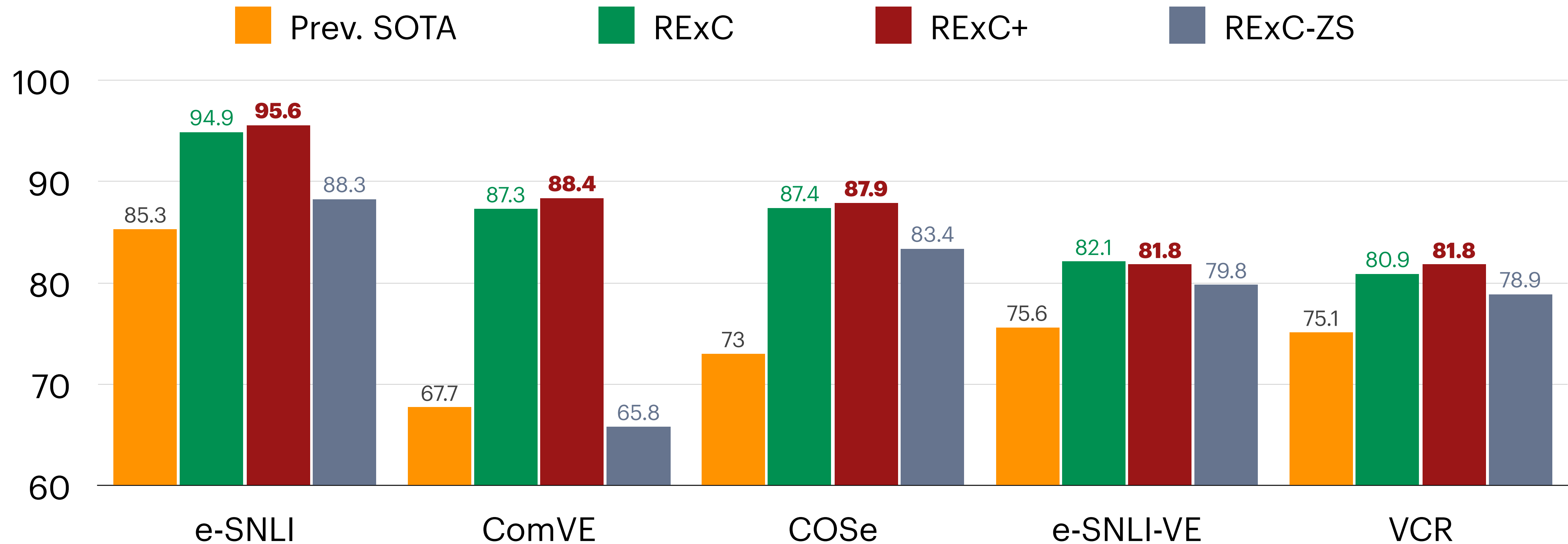


# How Rationale, Knowledge Help



Rationale and Selected Knowledge **individually contribute** to performance

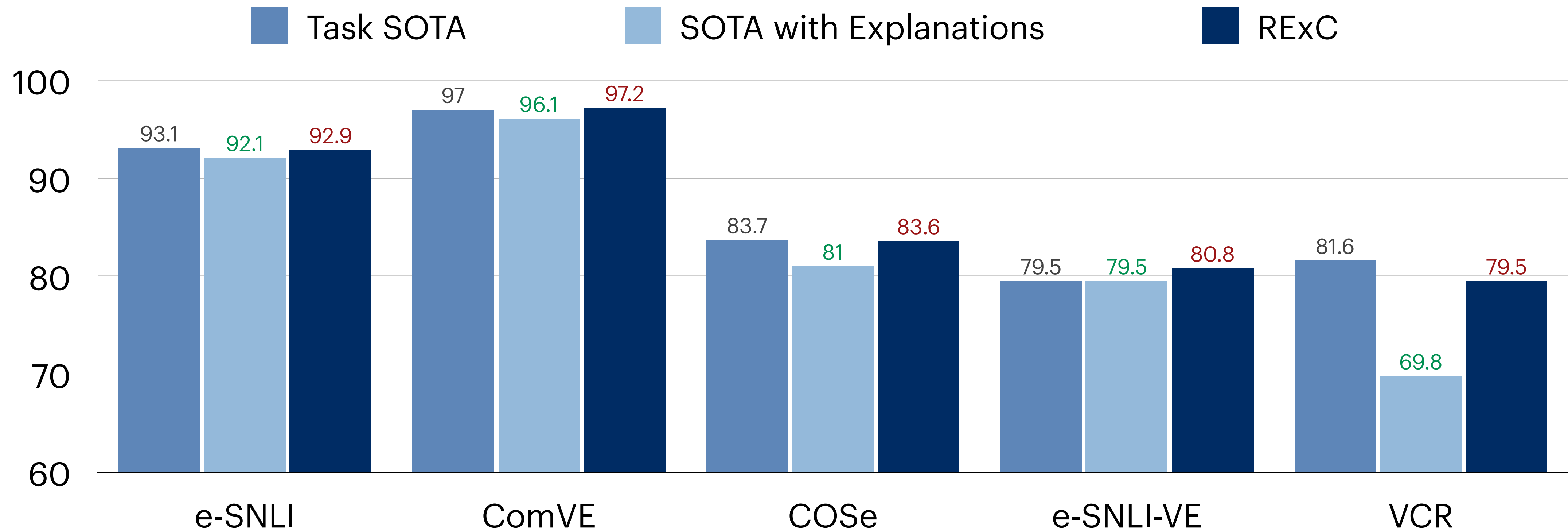
# Human Evaluation via e-ViL Scores



All RExC versions are **highly rated** by human users

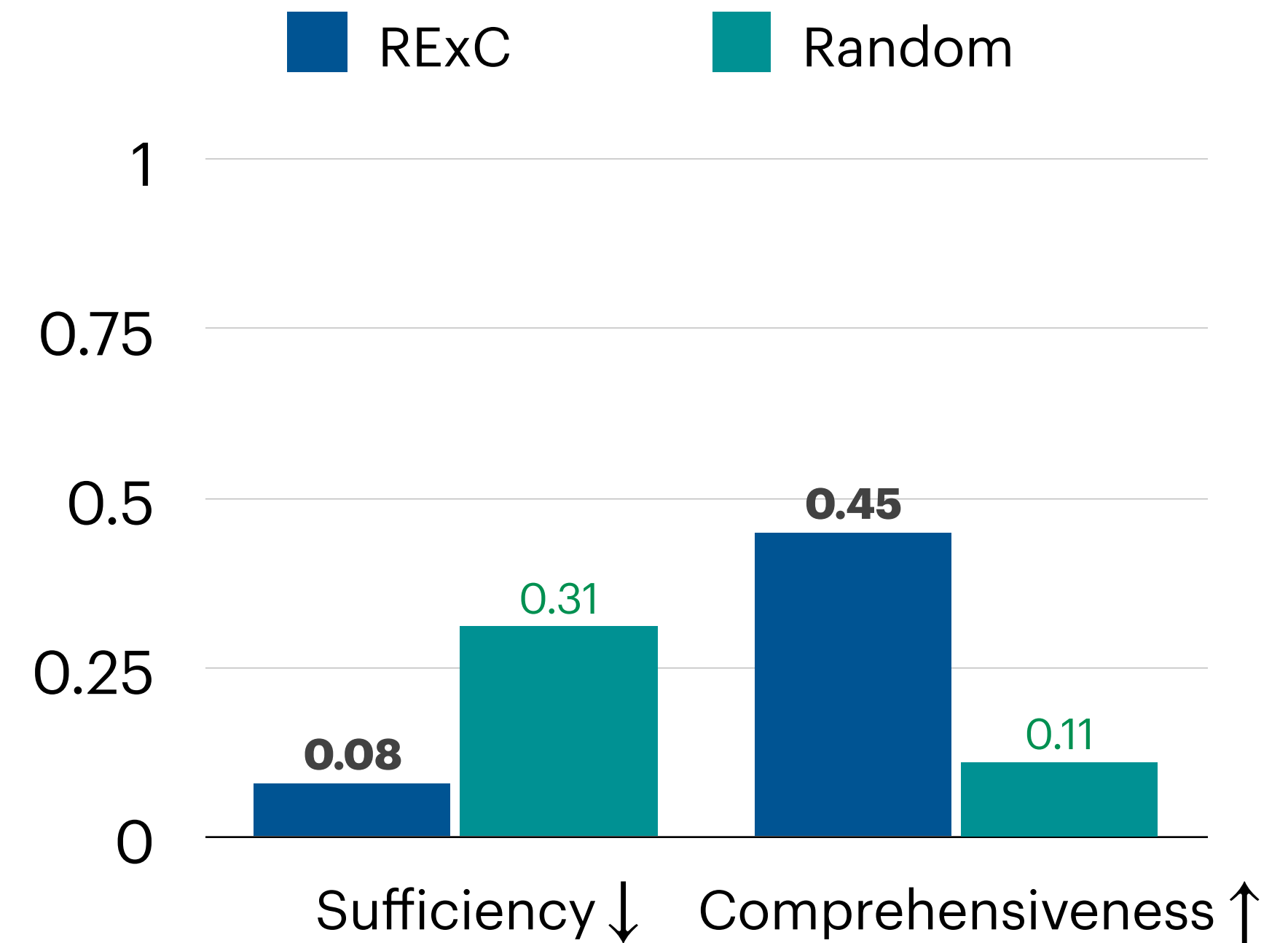
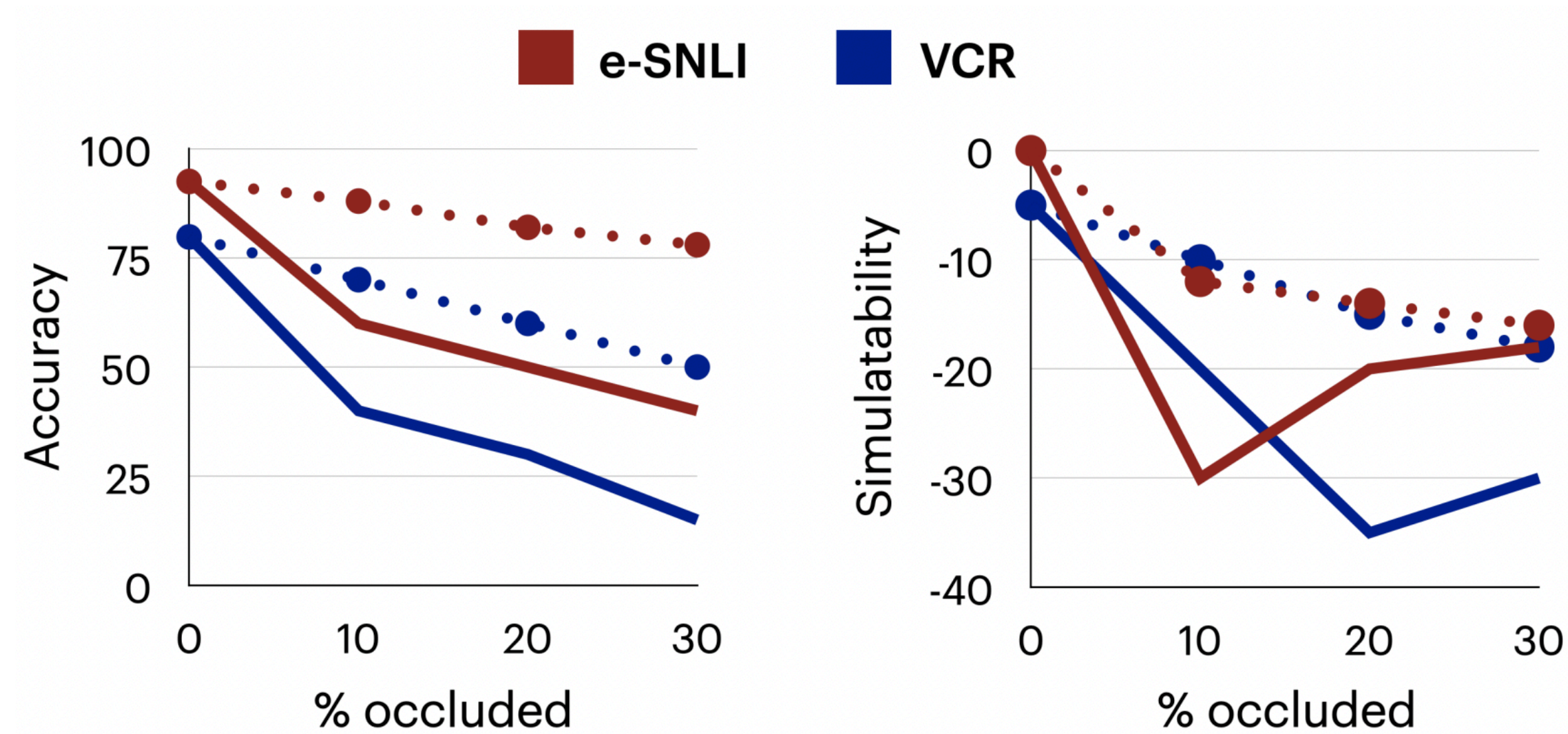


# RExC closing Performance-Explainability Gap



RExC is **task SOTA** for model **with explanations**,  
often outperforms other SOTA

# Faithful NLEs

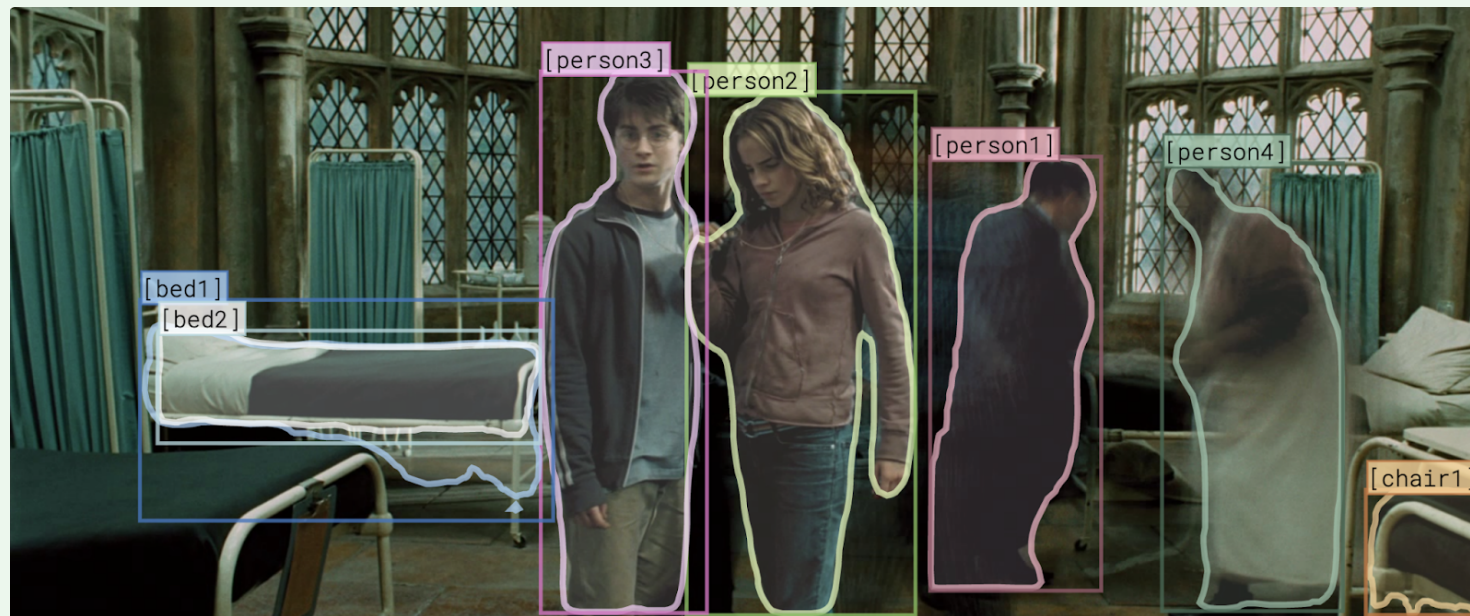


NLEs from RExC are **faithful** as both NLEs and task outputs are highly correlated  
Rationales are **sufficient** and **comprehensive**



# Summary

Q: Where are [person2] and [person3]?



A: They are in a hospital room

NLE: There are hospital beds and nurses in the room

Rationale:



Selected Knowledge:

Hospital room has hospital beds  
Hospital has nurses

- A **self-rationalizing** framework capable of producing both NLEs and rationales
- Generated NLEs are **grounded in background knowledge** obtained from rationales
- RExC **achieves SOTA** for NLE, Rationale quality, task performance
- RExC generated **explanations are faithful**, sufficient, and comprehensive

## Thanks!

Come at the **poster session 1**  
**today, Tue Jul 19**

**06:30 PM -- 08:30 PM (EDT) @ Hall E**