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

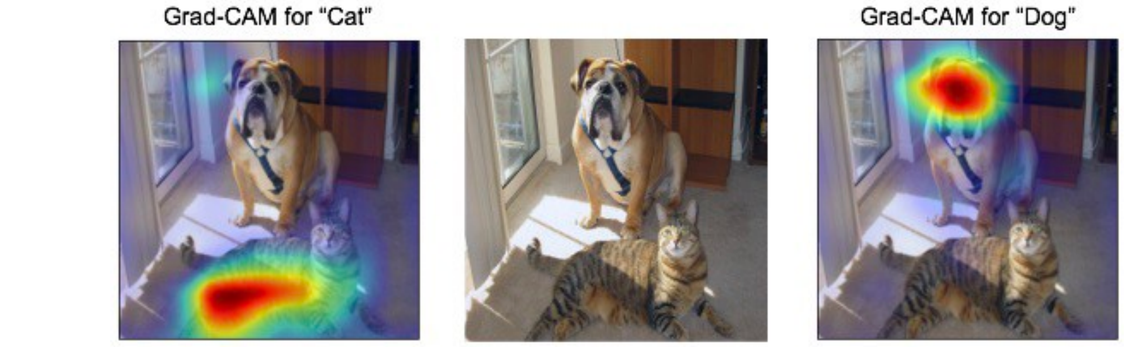
Bodhisattwa Prasad Majumder, Oana-Maria Camburu, Thomas Lukasiewicz, Julian McAuley
UC San Diego, University of Oxford, TU Wien Informatics

ICML 2022

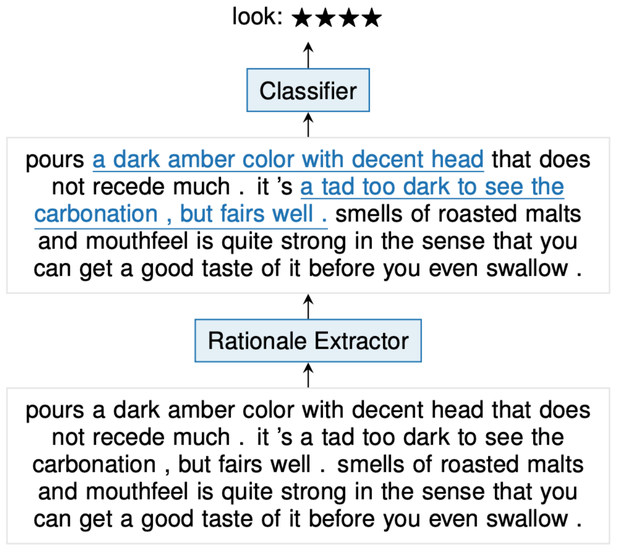
ICML
International Conference
On Machine Learning




User Experience with AI Explanations

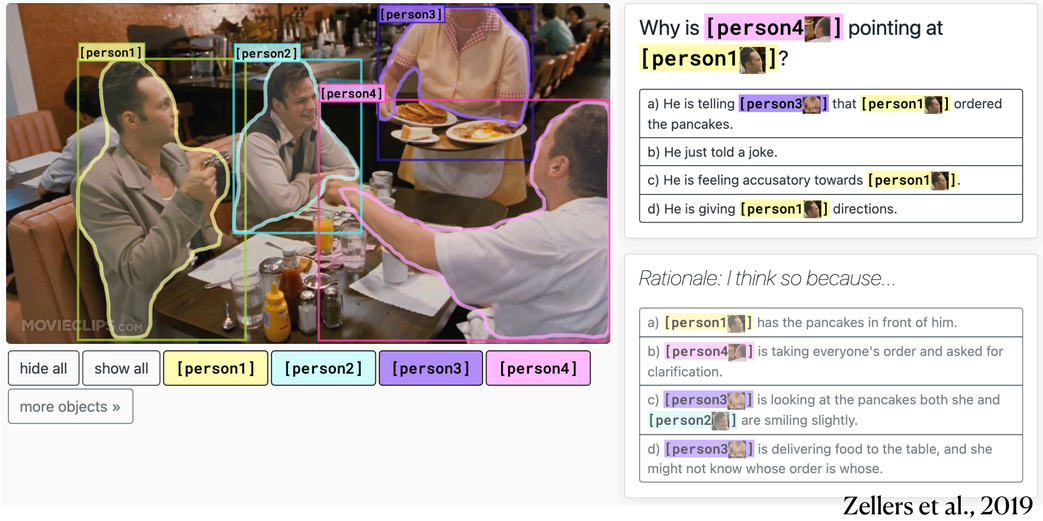


Selvaraju et al., 2019



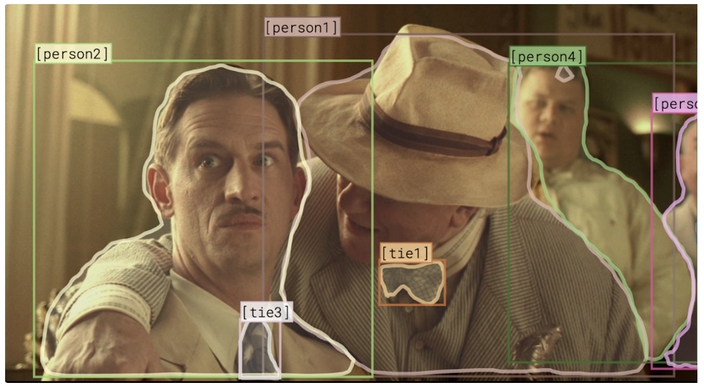
Bastings et al., 2020

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



Zellers et al., 2019

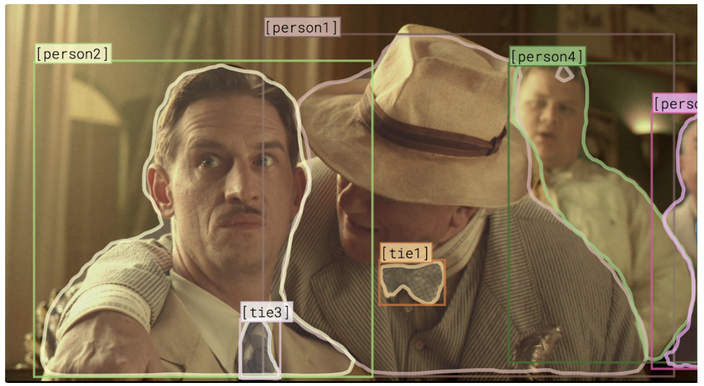
Rich Representation of Explanations



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

A: He's concerned and a little upset

extractive



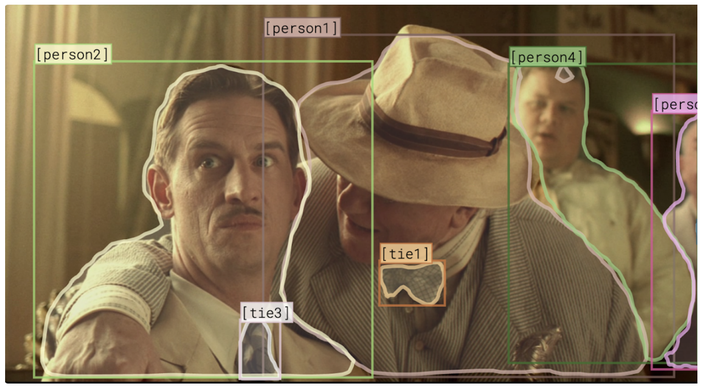
Q: how does `[person2]` feel about what `[person1]` is telling him?

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

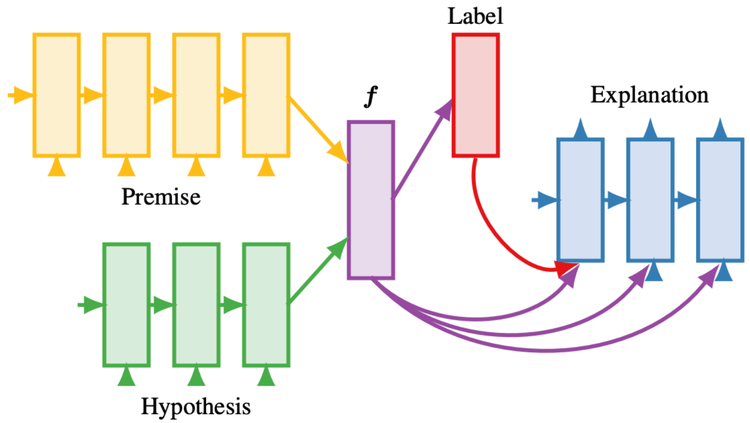
People are riding bikes

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

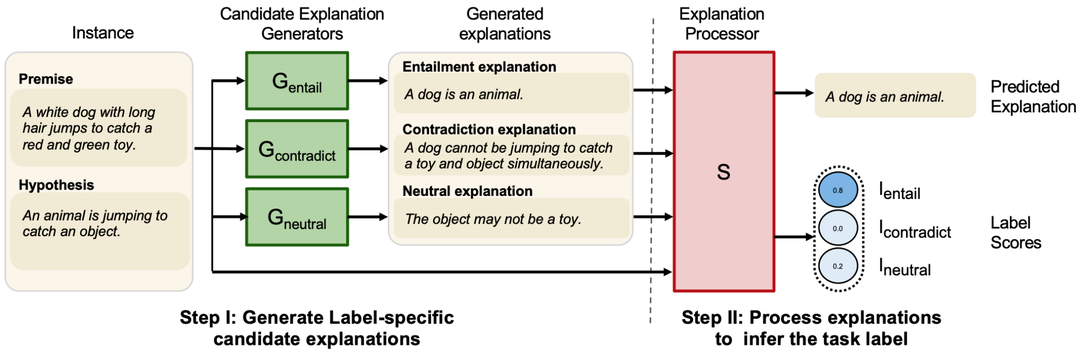
label
entailment

Competing in a bicycle race requires men riding bikes

Previous Works

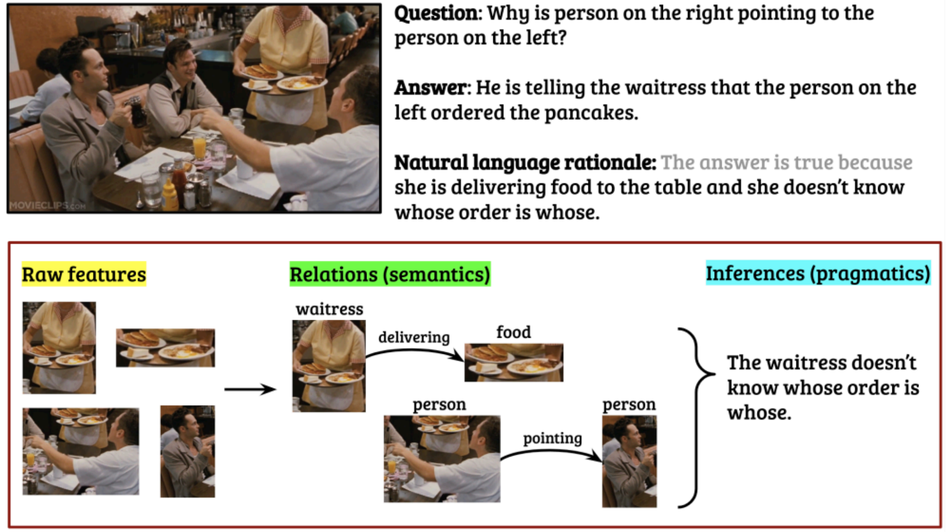


predict-then-explain (Camburu et al., 2018)



generate label-specific explanations, then choose the correct one

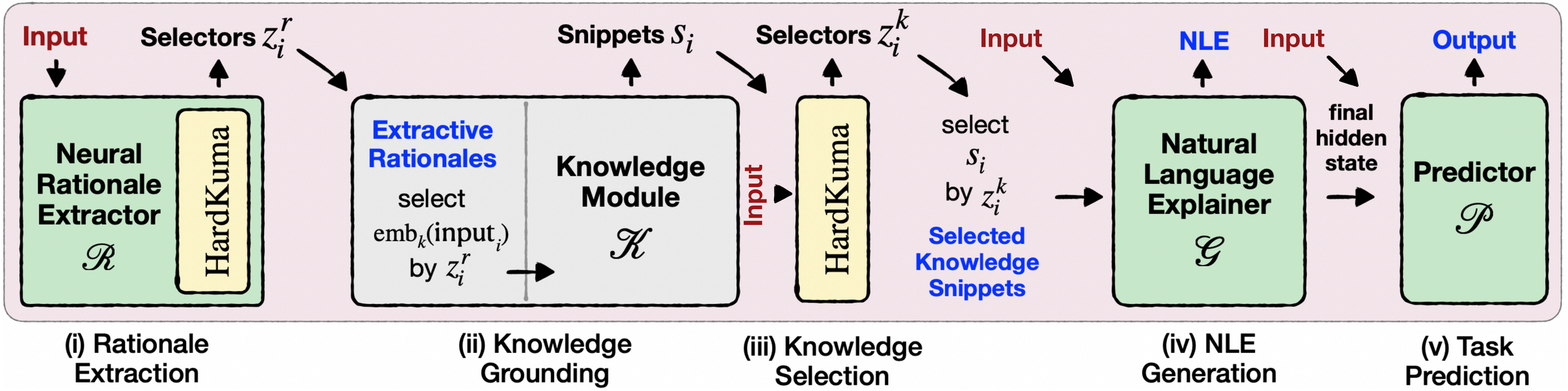
(Kumar et al., 2018)



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

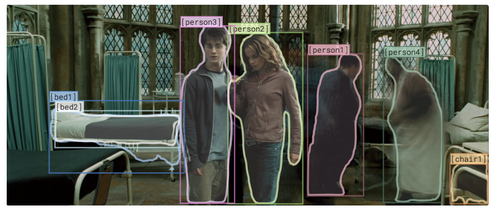
Q: Where does a wild bird usually live?

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

label: *sky*

Vision Language Tasks

Visual Entailment



Hypothesis: Some tennis players pose

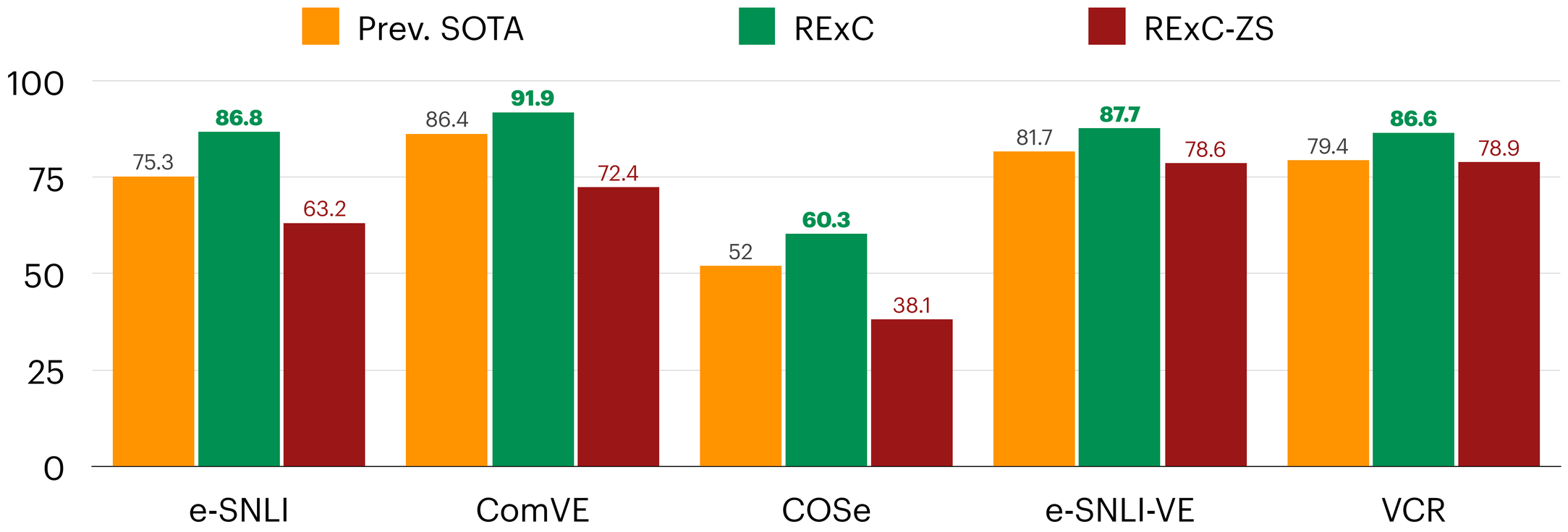
label: *entailment*

Visual Commonsense Reasoning

Q: What is the place?

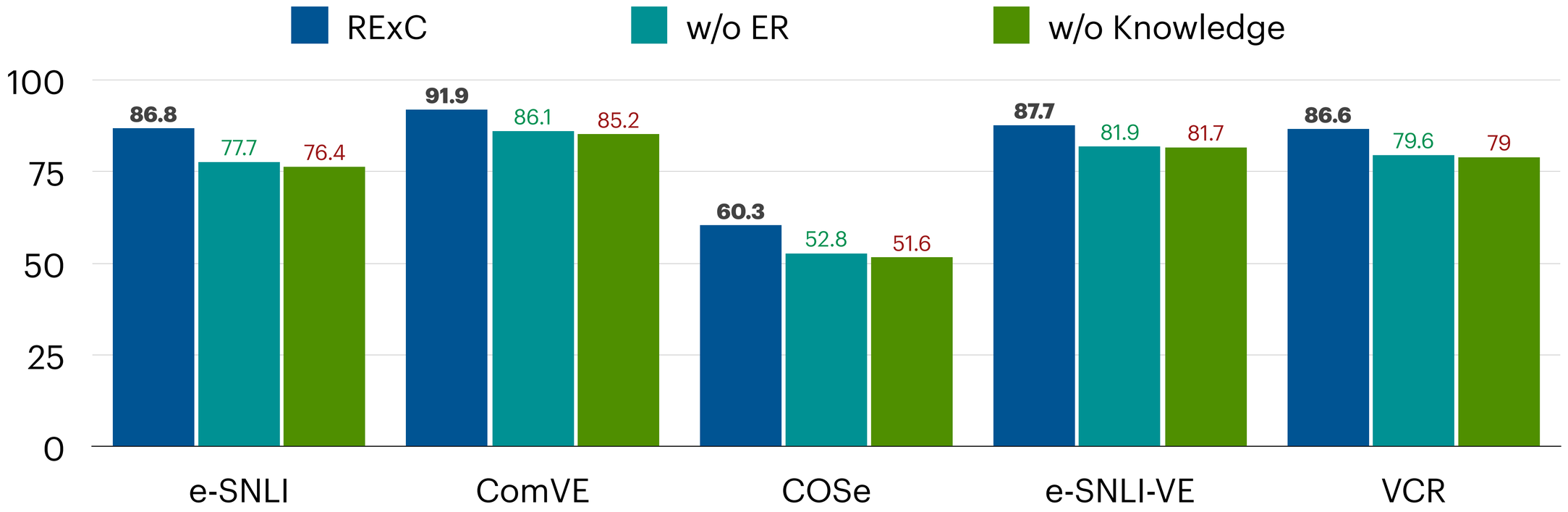
label: *They are in a hospital room*

Plausibility via BertScore



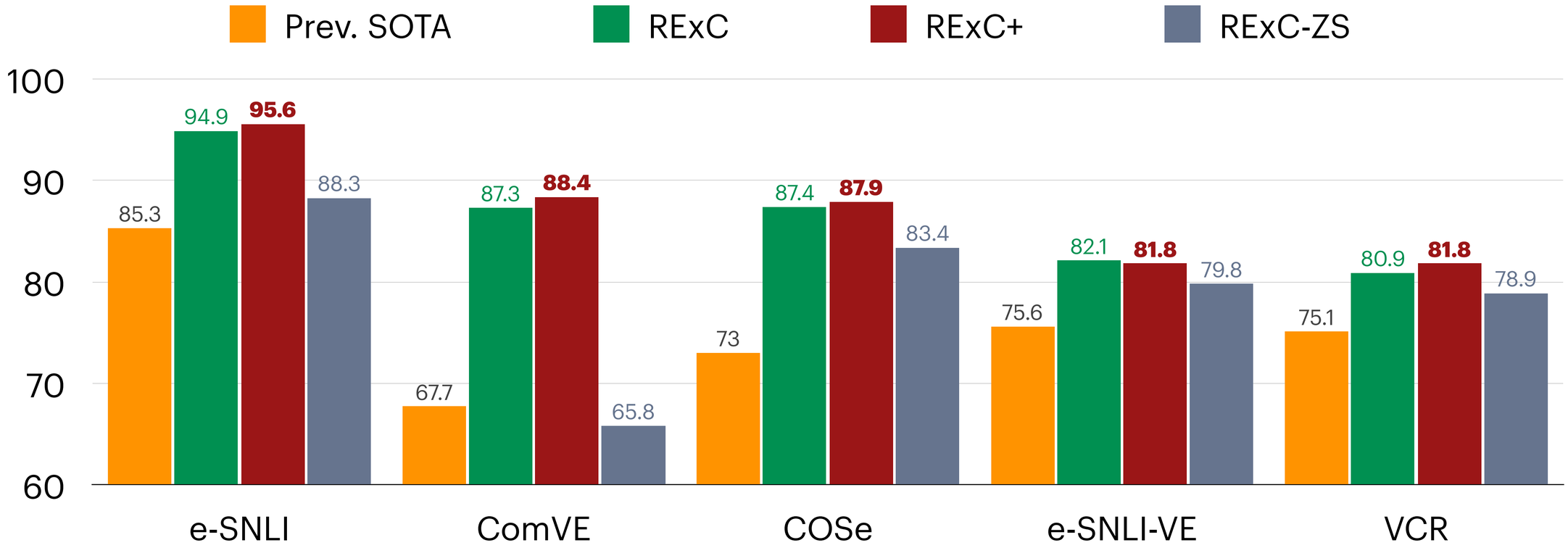
RExC **outperforms** all previous SOTA for NLE quality/plausibility

How Rationale, Knowledge Help



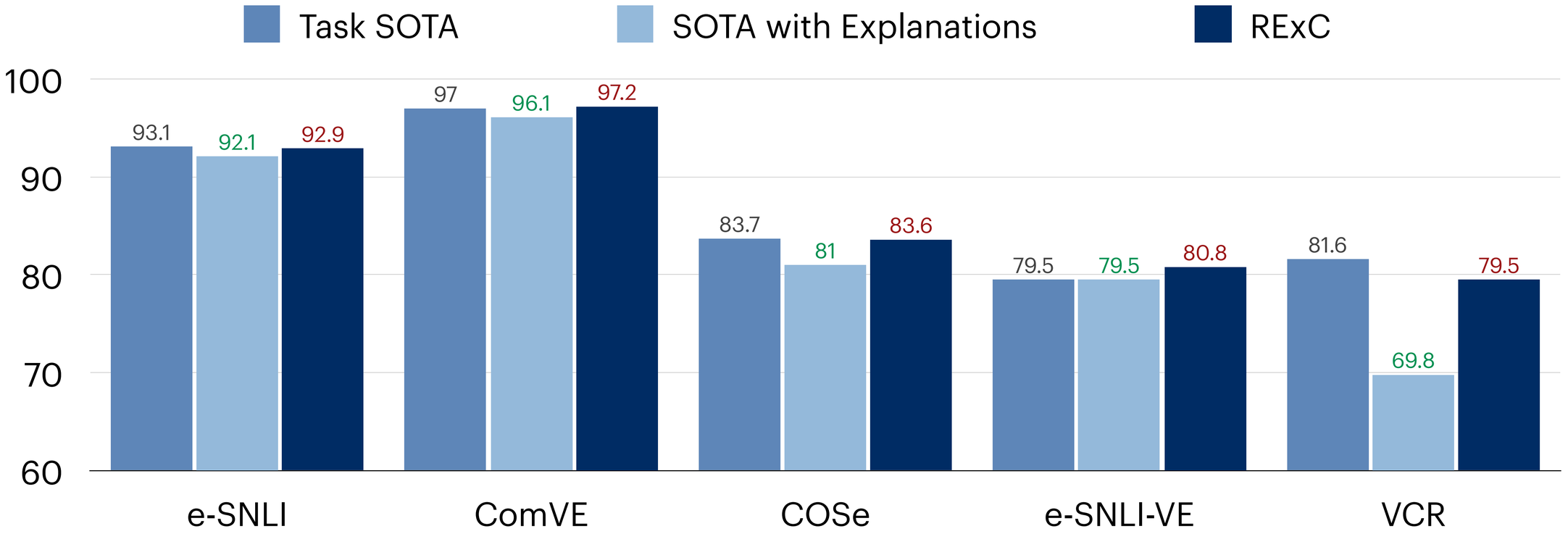
Rationale and Selected Knowledge **individually contribute** to performance

Human Evaluation via e-ViL Scores



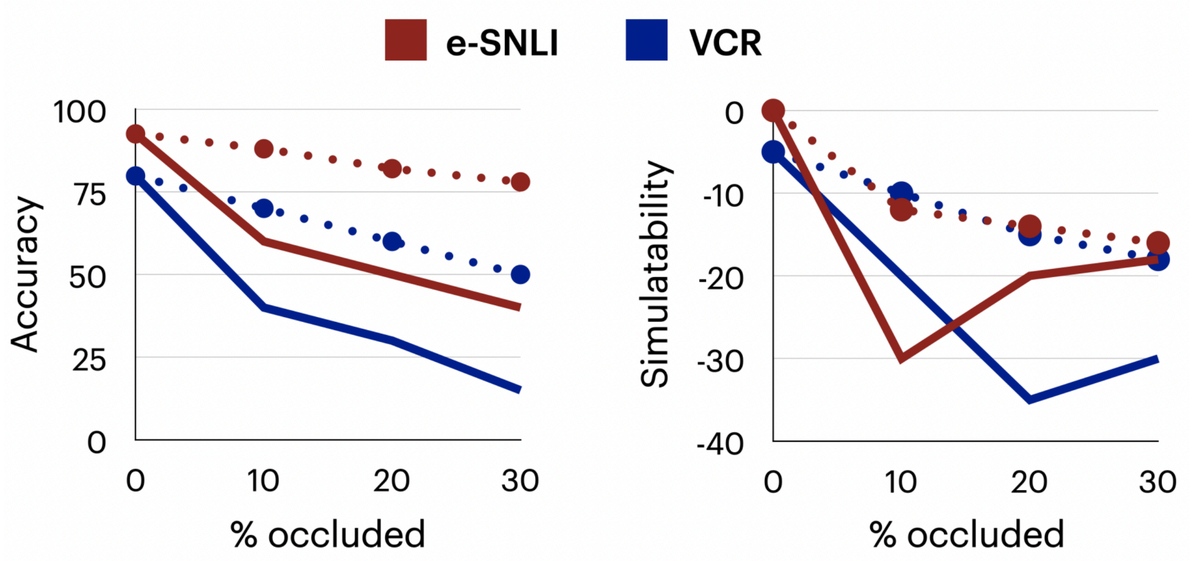
All RExC versions are **highly rated** by human users

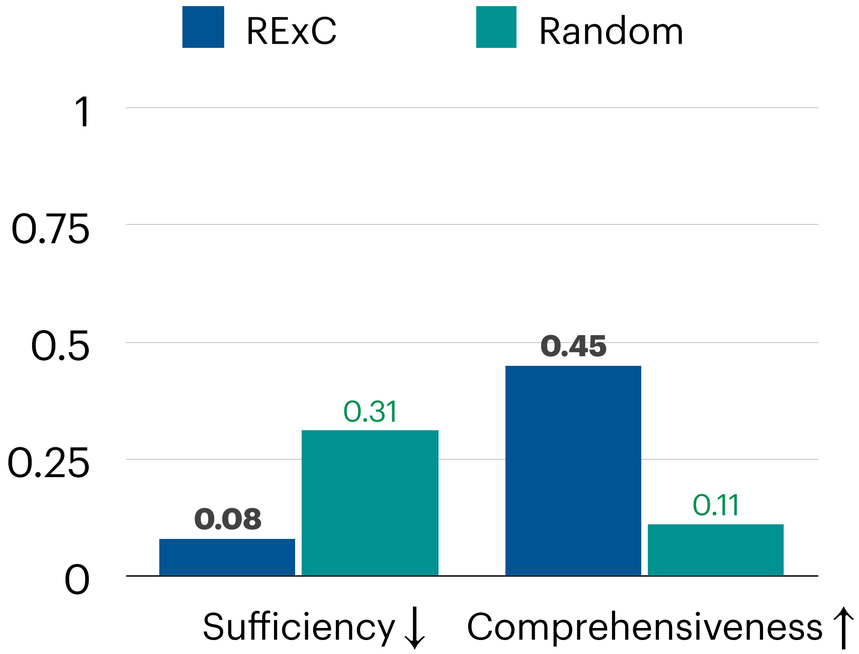
RExC closing Performance-Explainability Gap



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

Faithful NLEs



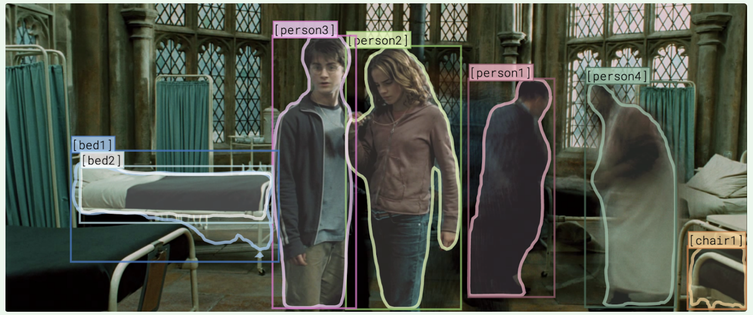


NLEs from RExC are **faithful** as both NLEs and task outputs are highly correlated

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