







REX: Reasoning-Aware and Grounded Explanation

Problems with existing networks

- Most recent studies in visual reasoning are dedicated to improving the accuracy of predicted answers, and less attention is paid to explaining the rationales behind the decisions
- Models commonly take advantage of spurious data biases

No.	antecedant words	antecedant visual words	consequents
1	what,time,day		afternoon*
2	what,time,day		night*
3	what,time,clock,show		11:30*
4	what,time,year		fall*

How the paper tackle this problems

1. A new type of multi-modal explanations (a new dataset with 1,040,830 multi-modal explanations)
2. A novel explanation generation method that explicitly models the pairwise correspondence between words and regions of interest
3. Demonstration of the effectiveness of the new data and method under different settings

Example



Question: What is common to the comb and the heart?

Answer: color.

Reasoning Process:

Select (comb)

Select (heart)

Common (comb, heart)



Explanation: Because both



and

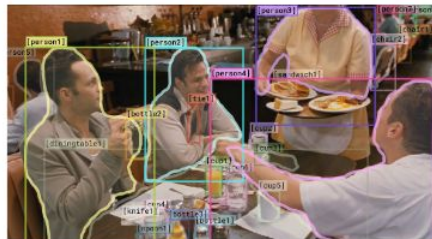


are **red**.

Figure 1. Illustration of our explanation that is derived from the reasoning process (with different reasoning steps color coded) and explicitly grounds key objects in the image.

Existing methods

- Multi-modal explanations for visual reasoning:
 - Explaining, elaborating, and enhancing your answers for visual questions
 - From Recognition to Cognition: Visual Commonsense Reasoning
- Generating multi-modal explanations:
 - Beyond VQA: Generating Multi-word Answers and Rationales to Visual Questions
 - Natural Language Rationales with Full-Stack Visual Reasoning: From Pixels to Semantic Frames to Commonsense Graphs



Why is [person4] pointing at [person1]?

- He is telling [person3] that [person1] ordered the pancakes.
 - He just told a joke.
 - He is feeling accusatory towards [person1].
 - He is giving [person1] directions.
- I chose a) 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 and both she and [person2] are smiling slightly.
 - [person3] is delivering food to the table, and she might not know whose order is whose.

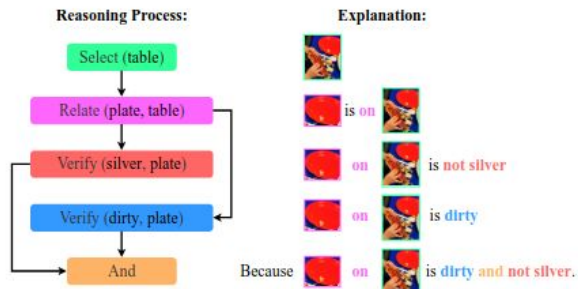


Data

This method considers evidence from both textual and visual modalities in an integral manner and couples words with image regions (i.e., for visual objects, their grounded regions instead of object names are considered in the explanations).



Question: Is the plate on the table both dirty and silver?
Answer: no.

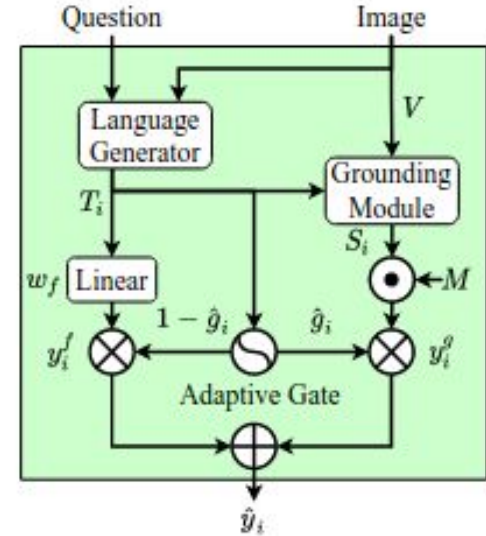


Operation	Semantic
Select	Selecting a specific category of objects.
Exist	Examining the existence of a specific type of objects.
Filter	Selecting the targeted objects by looking for a specific attribute.
Query	Retrieving the value of a attribute from the selected objects.
Verify	Examining if the targeted objects have a given attribute.
Common	Finding the common attributes among a set of objects.
Same	Examining if two groups of objects have the same attribute.
Different	Examining if two groups of objects have different attributes.
Compare	Comparing the values of an attribute between multiple objects.
Relate	Connecting different objects using their relationships.
And/Or	Logical operations that combine results of previous operations.

Table 1. Atomic operations to represent the reasoning process.

Model

- Existing explanation generation methods model textual and visual explanations with separate processes
- A novel explanation generation model couples related components across the two modalities and generates the explanation based on their relationships.



Comparative results

	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE	Grounding	GQA-val	GQA-test	OOD-val	OOD-test
VisualBert [21]	-	-	-	-	-	-	64.14	56.41	48.70	47.03
VisualBert-VQAE [22]	42.56	34.51	73.59	358.20	40.39	31.29	65.19	57.24	49.20	46.28
VisualBert-EXP [40]	42.45	34.46	73.51	357.10	40.35	33.52	65.17	56.92	49.43	47.69
VisualBert-REX	54.59	39.22	78.56	464.20	46.80	67.95	66.16	57.77	50.26	48.26

Comparative results on explanation generation and question answering. GQA- and OOD- denote results on GQA and GQA-OOD. Best results are highlighted in bold.

Comparative results

		1%		5%		10%	
		GQA	OOD	GQA	OOD	GQA	OOD
VQA-only	VisualBert	41.41	27.11	48.53	33.78	51.79	37.83
Multi-task learning	VisualBert-EXP	41.70	27.09	49.36	34.50	52.83	38.33
	VisualBert-REX	40.42	23.95	50.30	35.69	53.90	40.08
Self-supervised learning	VisualBert	45.06	30.62	52.12	38.68	54.74	40.12
Transfer learning	VisualBert-EXP	51.32	35.26	56.34	41.20	57.65	43.15
	VisualBert-REX	57.07	40.03	61.28	45.02	61.90	45.98

Comparative results for models trained using different proportions of answer annotations. Results are reported on the balanced validation set of GQA and the validation set of GQA-OOD. Best results are highlighted in bold.

Is the knowledge learned from the explanations transferable to question answering?

- Knowledge transferred from the explanations plays a key role in question answering.
- Visual grounding is important for transferring knowledge.

VisualBert-VQAE



Question: Are both the car and the horse the same color?

Answer: yes.



Predicted Answer: yes.

Explanation: Because #1 is black and #2 is white.

VisualBert-EXP



Predicted Answer: no.

Explanation: Because #1 is black and #1 is white.

VisualBert-REX



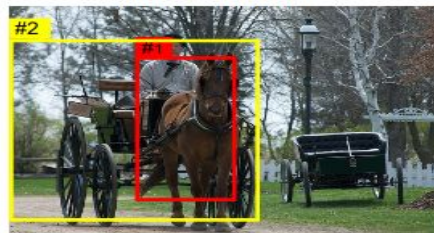
Predicted Answer: yes.

Explanation: Because #1 and #2 are both black.



Question: Does the carriage on the grass appear to be black and small?

Answer: yes.



Predicted Answer: no.

Explanation: Because there is #1 on #2 that is black and not small.



Predicted Answer: no.

Explanation: Because there is #1 on #2 that is black and not small.



Predicted Answer: yes.

Explanation: Because there is #1 on #2 that is black and small.



Question: What color do you think the balloon to the right of the giraffes is?

Answer: white.



Predicted Answer: white.

Explanation: Because #1 to the right of #2 is white.



Predicted Answer: white.

Explanation: Because #1 to the right of #1 is white.



Predicted Answer: white.

Explanation: Because #1 to the right of #2 is white.

How do different visual skills affect answer correctness?

- Recognition of attributes is important for answering correctly
- Attributes do not contribute equally to answer correctness.

Broader Impact

- **important step toward trustworthy AI**
- **Could be very important for example in medicine or finance**



Question: Are both plates and forks in the picture?

Answer: yes

Explanation: Because there are # 1 and #2.



Question: Is there any surfboard to right of the man the people are standing by?

Answer: yes

Explanation: Because # 1 is to the right of #2 standing near #3.



Question: Do you see a chair to the left of pillow?

Answer: no

Explanation: Because there is no chair to the left of #1.



Question: What do the window and the bed have in common?

Answer: shape

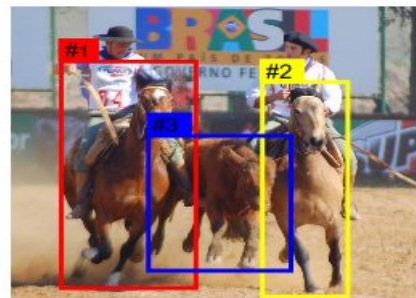
Explanation: Because both #1 and #2 are round.



Question: What do the end table and the frame have in common?

Answer: material

Explanation: Because both #1 and #2 are wood.



Question: Are these animals of different types?

Answer: yes

Explanation: Because #1 and #2 are horse, #3 is bull.



Thank you for your
attention