LRTA: A Transparent Neural-Symbolic Reasoning Framework with Modular Supervision for Visual Question Answering

Weixin Liang[1], Feiyang Niu[2], Aishwarya Reganti[2], Govind Thattai[2], Gokhan Tur[2]

[1]Department of Electrical Engineering, Stanford University, US

[2]Alexa AI, Amazon, US

Multimodal Question Answering

1

- **Problem**: answer free-form questions by reasoning about presented images
- **Dataset**: GQA
  - 113,018 images & 1.5M questions
  - 1702 object classes

Scene graph

```
{
  "2407890": {
    "objects": {
      "271881": {
        "name": "tray",
        {"x": 0, "y": 288, "w": 448, "h": 192},
        "attributes": ["red", "plastic"],
        "relations": {
          "32452": {
            "name": "to the left of",
            "object": "275312"
          },
          ...
        }
      },
      ...
    }
  }
}
```

Question semantics

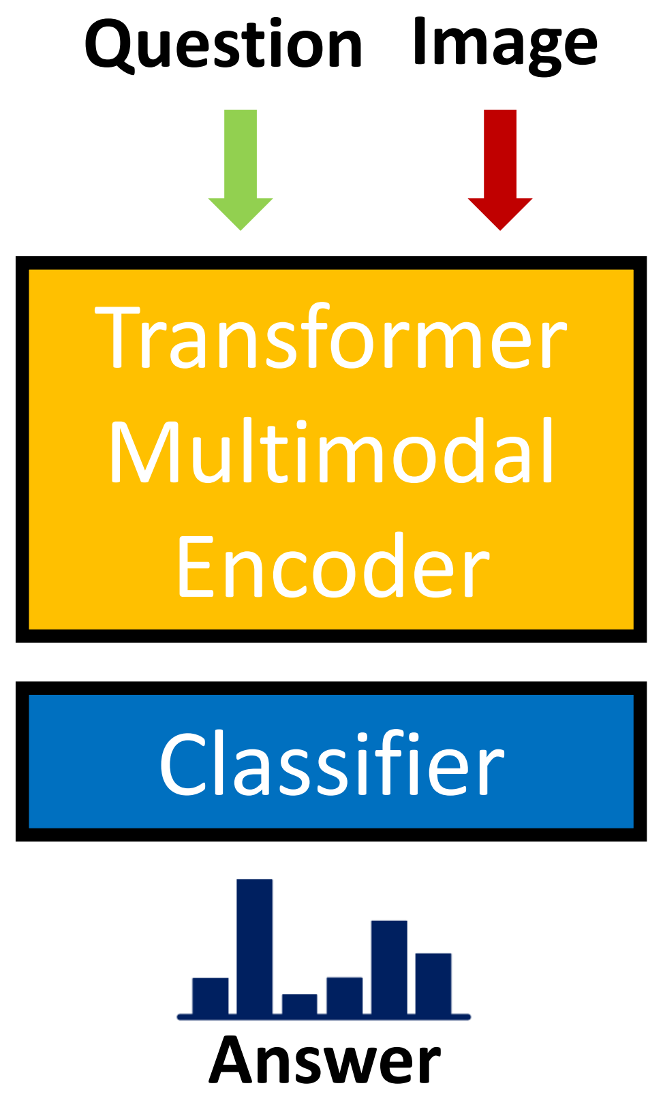
```
{
  "1238592": {
    "imageId": "2407890",
    "question": "Is there any red object to the left of the girl who is holding a hamburger?",
    "answer": "yes",
    "fullAnswer": "Yes, there",
    "isBalanced": true,
    ...
    "semanticStr": "select: hamburger (279472) -> relate: holding, subject, girl (271881)
                                              -> related: to the left of, subject, tray (278452)
                                              -> filter: red
                                              -> exist: ?"
  }
}
```

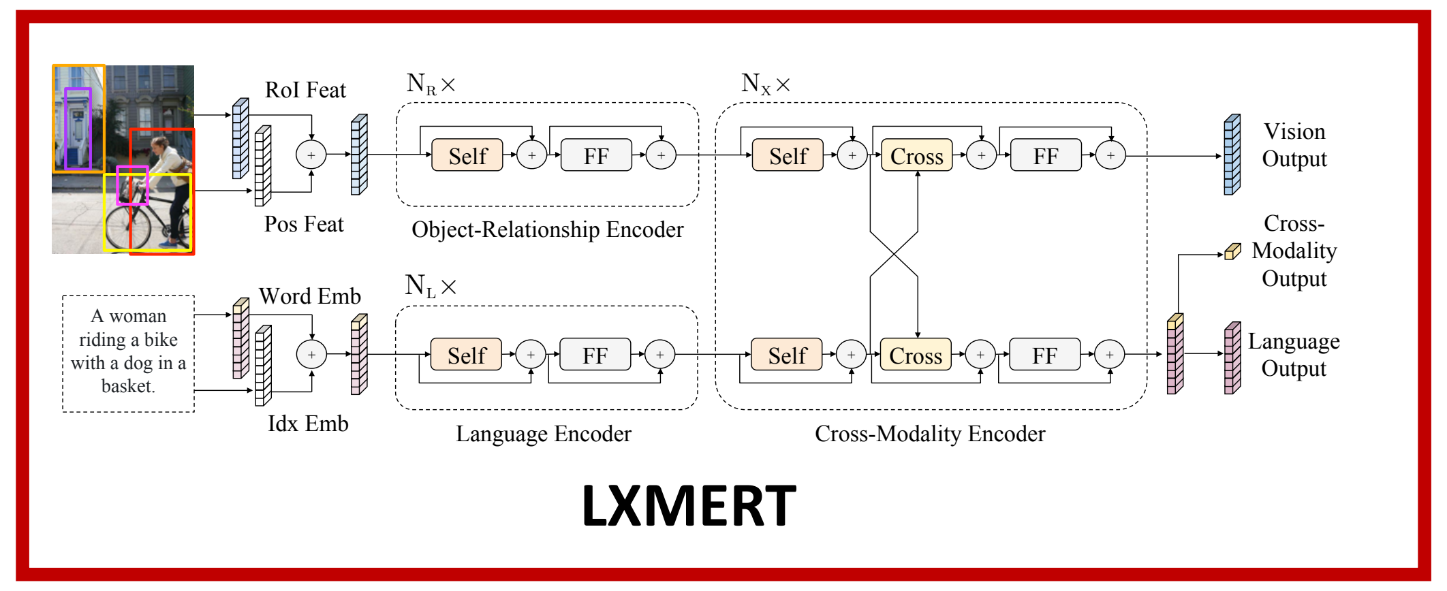
Question:
Is there any red object to the left of the small girl who is holding a hamburger?

Short Answer: Yes
Long Answer: Yes, there is a red tray

Motivation

2





LXMERT

Our Main Baseline in this talk

Limitations

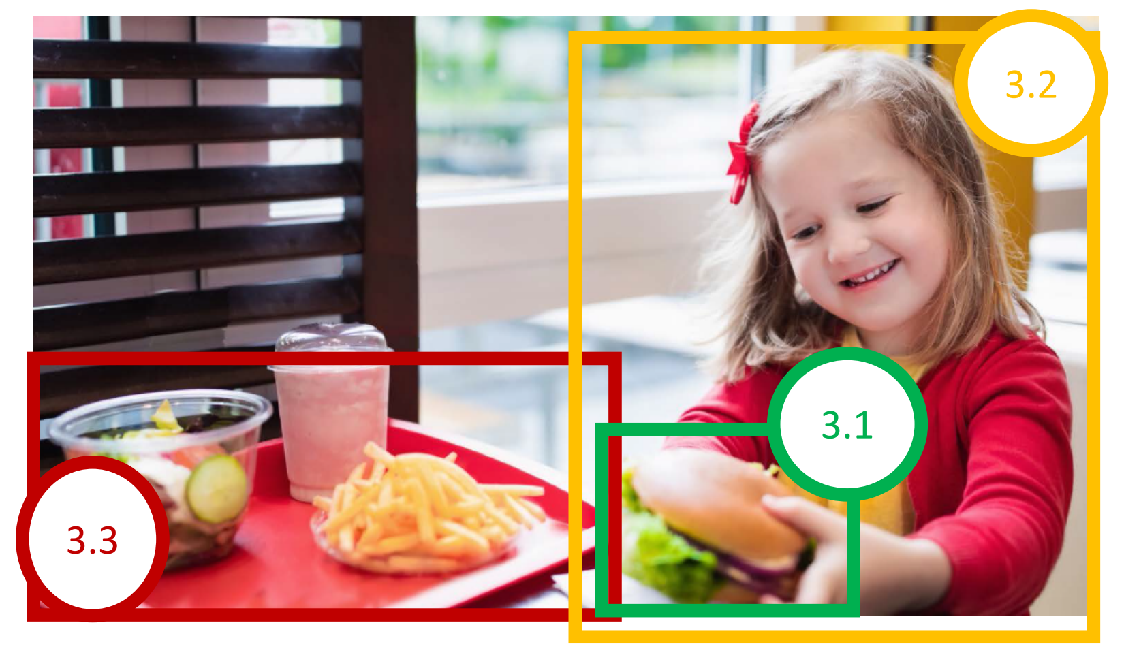
- LXMERT not really understands the question
  - Mask out relationship words (e.g. “to the left of”) in the questions: 59.7% → 55.5% (less than 5% drop)[1]
- A “black-box” neural encoder without human readable justification

[1] Sanjay Subramanian, Sameer Singh, and Matt Gardner. 2019. Analyzing compositionality of visual question answering. Visually Grounded Interaction and Language Workshop.

LRTA: A More Explainable Approach

3

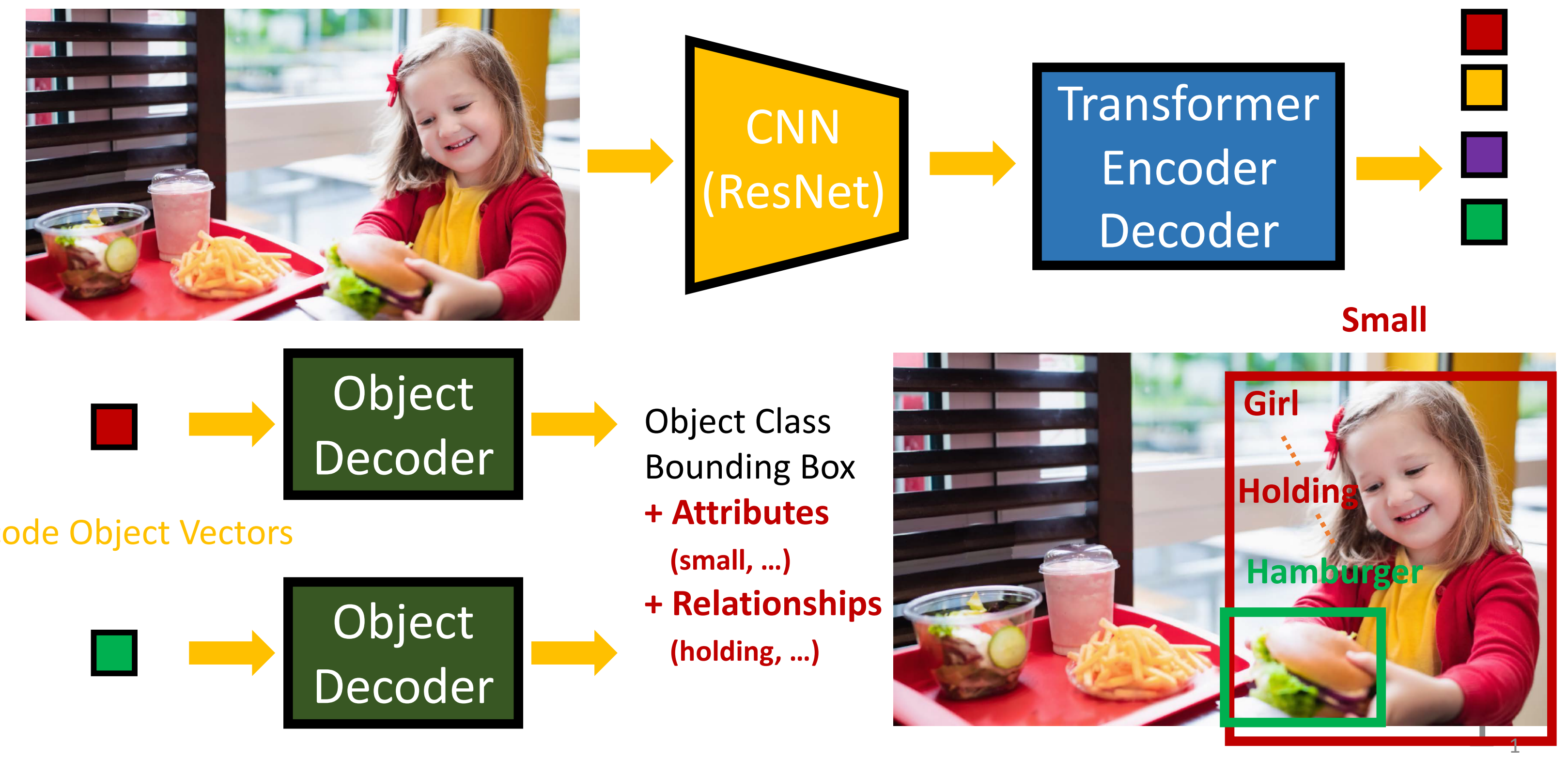
- **Question:** Is there any red object to the left of the small girl who is holding a hamburger?
- **LRTA:** Solving the problem step-by-step like humans
  - (1) Look at the image
  - (2) Read the question
  - (3) Think (Multi-Step)
    - (3.1) hamburger
    - (3.2) small girl
    - (3.3) tray
  - (4) Answer



Look: Scene Graph Generation

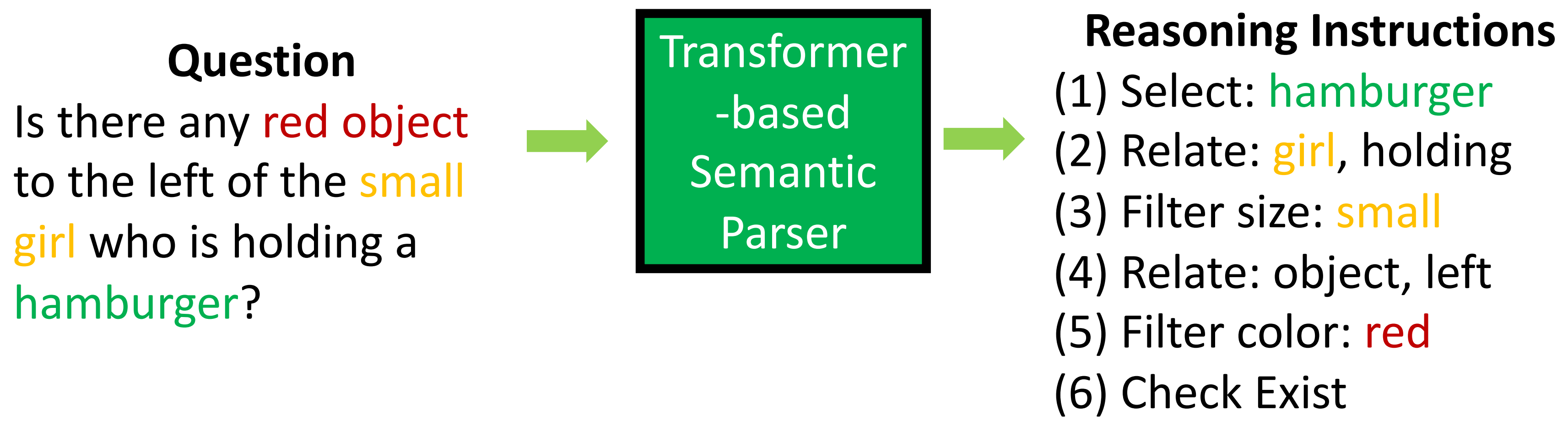
4

- Extended Facebook DETR: Object + Bounding Box + Attributes + Relationships



Read: Semantic Parser Module

5



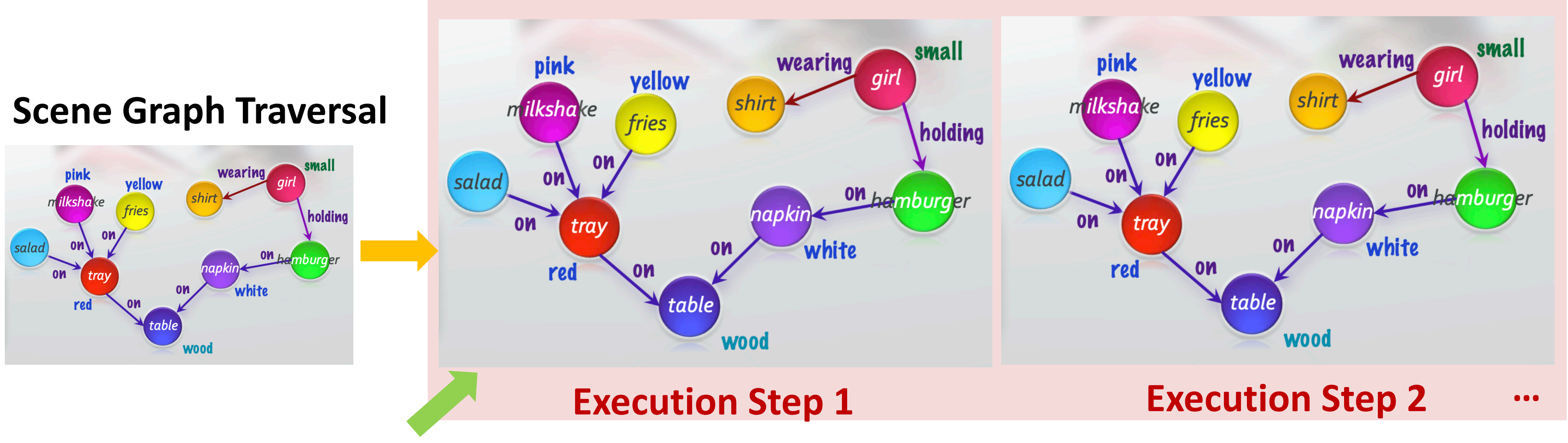
Question
Is there any red object to the left of the small girl who is holding a hamburger?

Reasoning Instructions
(1) Select: hamburger
(2) Relate: girl, holding
(3) Filter size: small
(4) Relate: object, left
(5) Filter color: red
(6) Check Exist

Reason: Neural Execution Module

6

Scene Graph Traversal

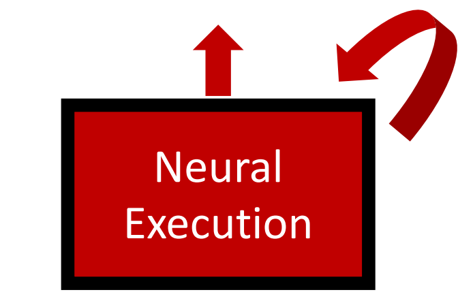


Execution Step 1

Execution Step 2 ...

Reasoning Instructions
(1) Select: hamburger
(2) Relate: girl, holding
(3) Filter size: small
...

Traversal node: Girl



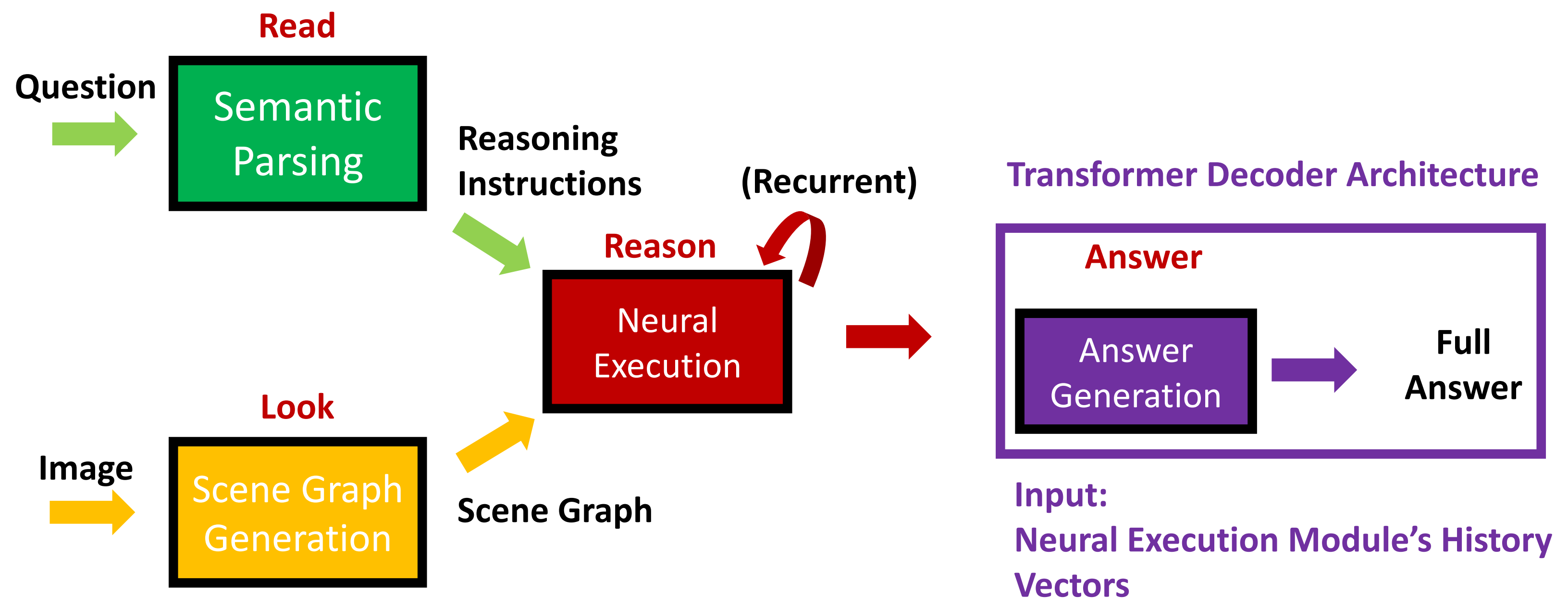
Instruction Vector
Relate: girl, holding

(1) Send to Each Node
- **History Vector**
- **Instruction Vector**
- Relate: girl, holding

(2) Each node queries its immediate neighbors and updates itself
(3) Aggregate all nodes
(4) Update history and output traversal node

Answer: Natural Language Generation

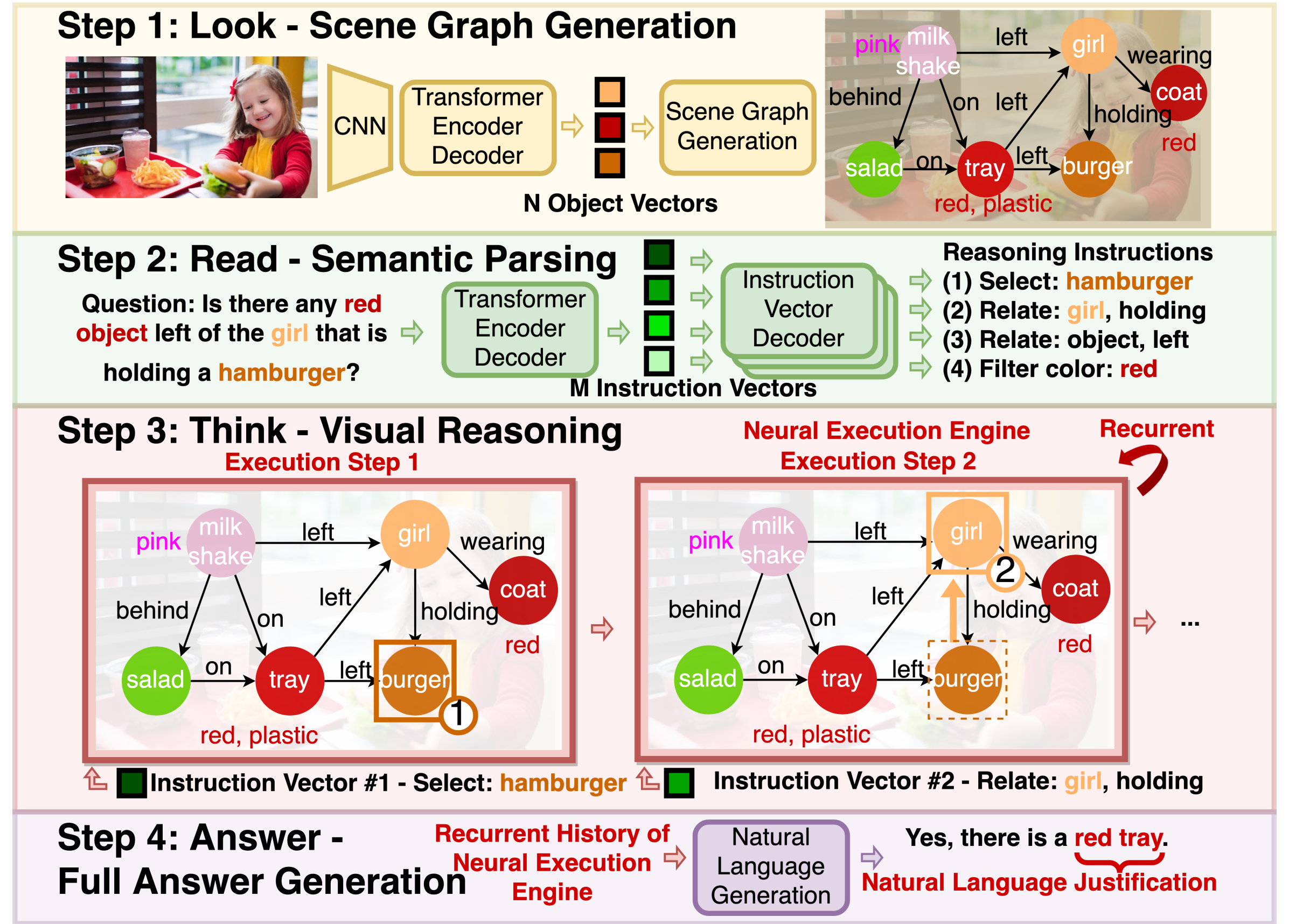
7



Input:
Neural Execution Module’s History Vectors

LRTA Overview

8


Step 1: Look - Scene Graph Generation
CNN
Transformer Encoder Decoder
Scene Graph Generation
N Object Vectors
pink
milk shake
left
girl
wearing
coat
behind
on
holding
red
salad
tray
burger
red, plastic
Step 2: Read - Semantic Parsing
Question: Is there any red object left of the girl that is holding a hamburger?
Transformer Encoder Decoder
Instruction Vector Decoder
M Instruction Vectors
Reasoning Instructions
(1) Select: hamburger
(2) Relate: girl, holding
(3) Relate: object, left
(4) Filter color: red
Step 3: Think - Visual Reasoning
Neural Execution Engine
Recurrent
Execution Step 1
Execution Step 2
1
2
...
Instruction Vector #1 - Select: hamburger
Instruction Vector #2 - Relate: girl, holding
Step 4: Answer - Full Answer Generation
Recurrent History of Neural Execution Engine
Natural Language Generation
Yes, there is a red tray.
Natural Language Justification

Experiments

9

Model	Long Acc	Short Acc
Human	-	89.30%
Bottom-up	-	49.74%
MAC	-	54.06%
LXMERT	28.00%	**56.20%**
LRTA	**43.10%**	**54.48%**

Table 1: End-to-end training experiment on testdev set

Model	Long Acc	Short Acc
LRTA trained w/ visual oracle		
Evaluated w/o attributes	67.79%	78.21%
Evaluated w/o relations	67.95%	75.47%
Evaluated w/o attributes & relations	50.15%	61.15%
Evaluated w/ visual oracle	**85.99%**	**93.10%**
LRTA trained w/ reading oracle		
Evaluated w/ reading oracle	55.45%	64.36%

Table 2: Validation study on valid set

Model	Short Acc Drop (from → to)
VB & PRPN masked	
LXMERT	19.43% (56.20% → 36.77%)
LRTA	**26.20%** (54.48% → 28.28%)
Attributes masked	
LXMERT	9.41% (56.20% → 46.79)
LRTA	**21.03%** (54.48% → 33.45)

Table 3: Perturbation analysis on testdev set. The larger drop the better

Conclusions

10

- **Contributions**
  - We propose LRTA, an end-to-end trainable, modular VQA framework facilitating explain-ability and enhanced error analysis as compared to contemporary black-box approaches.
  - We formulate VQA as a full answer generation problem to improve explainability and discourage superficial guess for answering the questions.
- **Future works**
  - Visual understanding poses as a bottleneck from our validation study and more model architectures should be explored and compared.
  - Scene graph data exhibit heavy long-tailed bias and an unbiased scene graph prediction needs to be explored, e.g. Tang et al 2020[1]

[1] K. Tang, Y. Niu, J. Huang, J. Shi, and H. Zhang. Unbiased scene graph generation from biased training. In2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR2020, Seattle, WA, USA, June 13-19, 2020, pages 3713–3722. IEEE, 2020.