Towards Cognition-Inspired Vision and Language Methods

Jing Yu

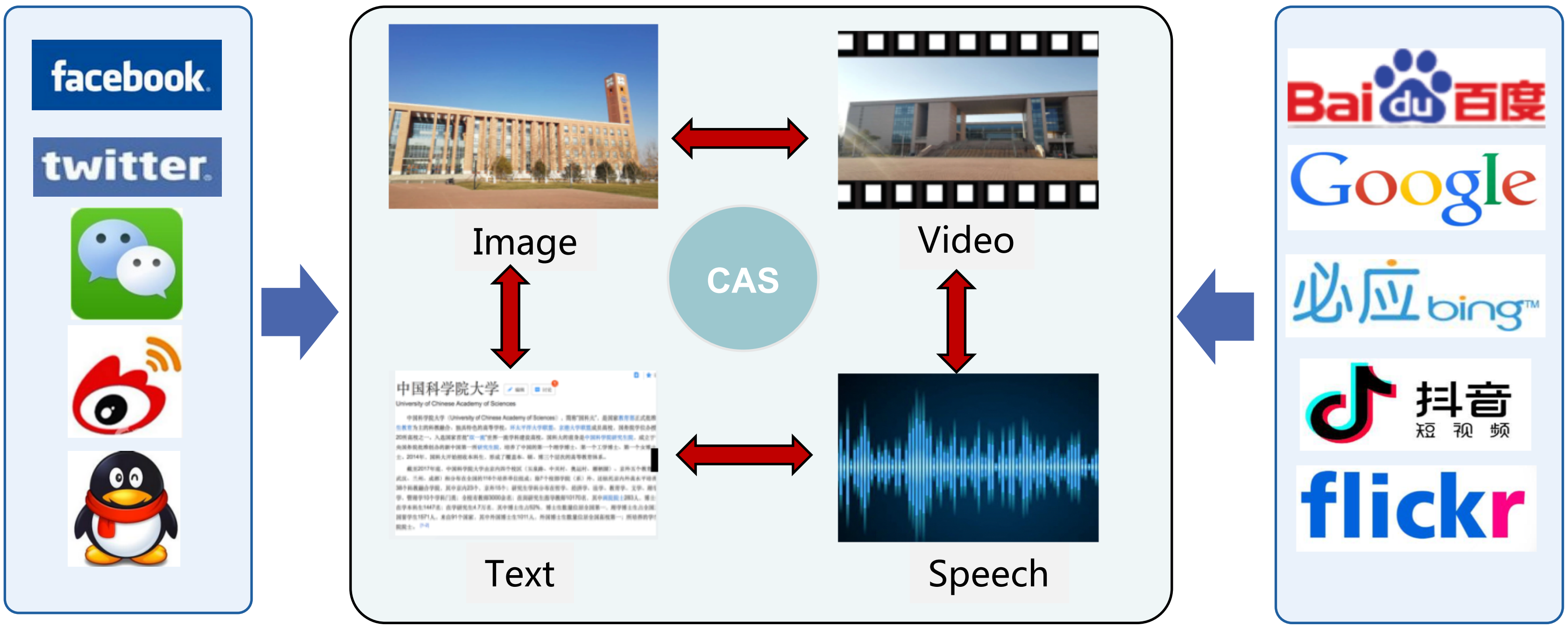
Institute of Information Engineering, CAS
University of Chinese Academy of Sciences

April 24th, 2021@CCF YOCSEF西安技术论坛

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

中国科学院大学
University of Chinese Academy of Sciences

Cross-Modal Intelligence



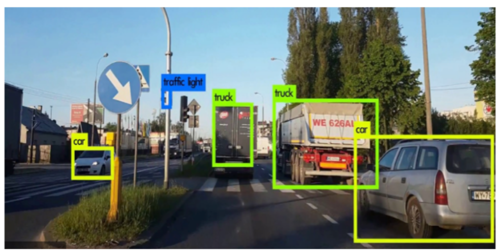
Vision-and-Language Tasks

CV + NLP = Vision-to-Language

CV

- Image Understanding
- Image Classification

- Object Detection



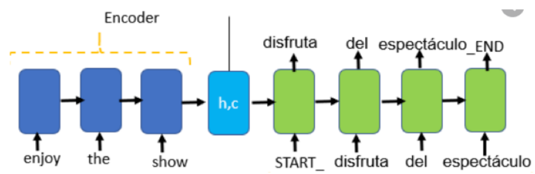
- Segmentation

- Object Counting
- Color Analysis
- …

\+

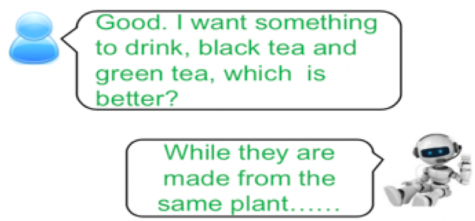
NLP

- Language Generation



\+

- Question Answering



\+

- Dialogue

User: how old are you?
Machine: I am three years old.

User: do you like to play piano?
Machine: Yes, I play piano.

=

Vision-to-Language

- Image Captioning

A zebra standing on top of a rocky field.

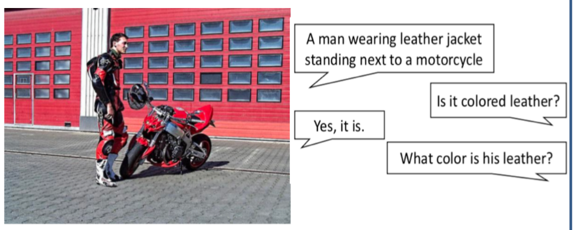
=

- Visual Question Answering

A1. Is the tray on top of the table black or light brown? light brown
A2. Are the napkin and the cup the same color? yes
A3. Is the small table both oval and wooden? yes

=

- Visual Dialogue



Three Stages in Vision-and-Language

Prof. Qi Wu's VL pyramid

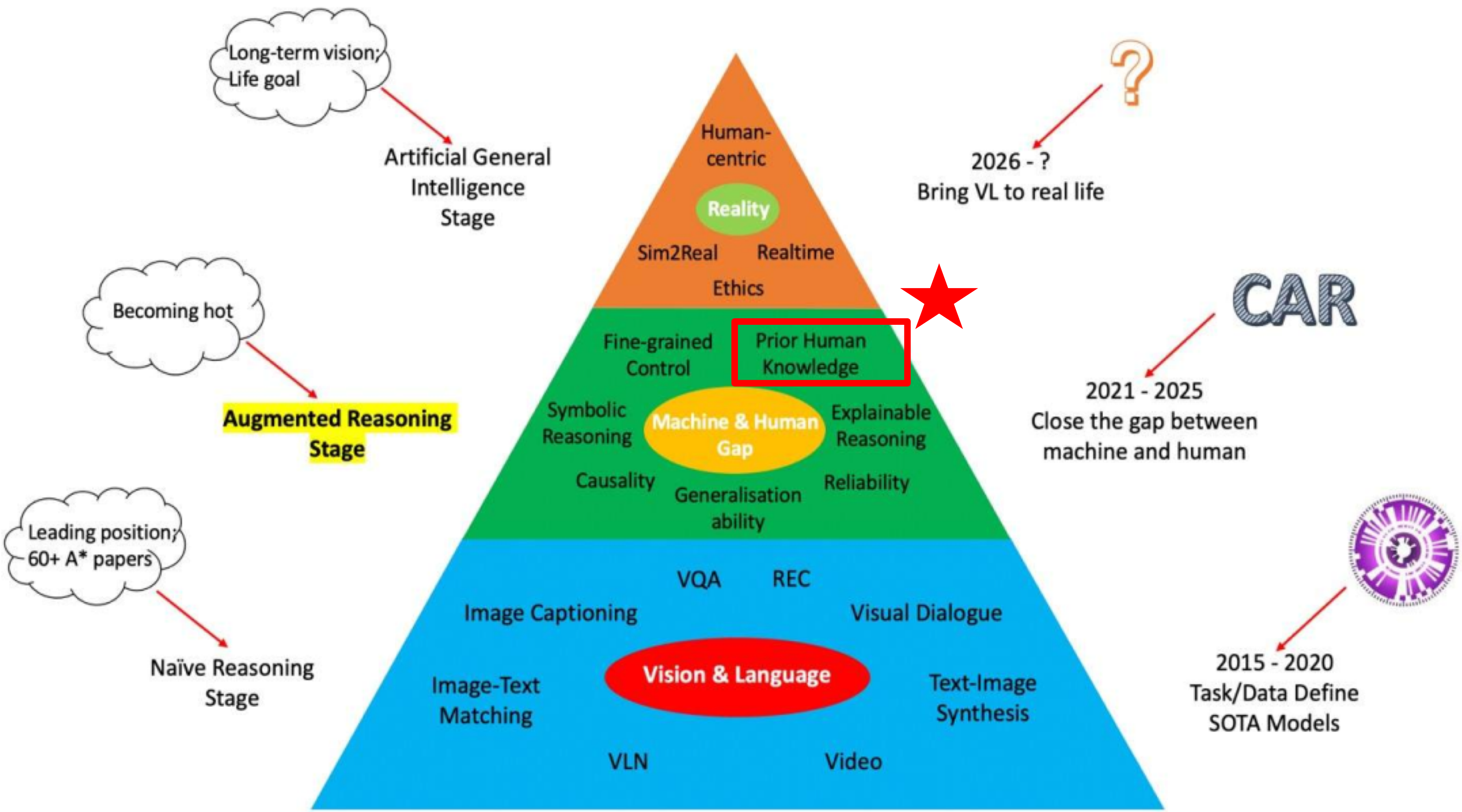
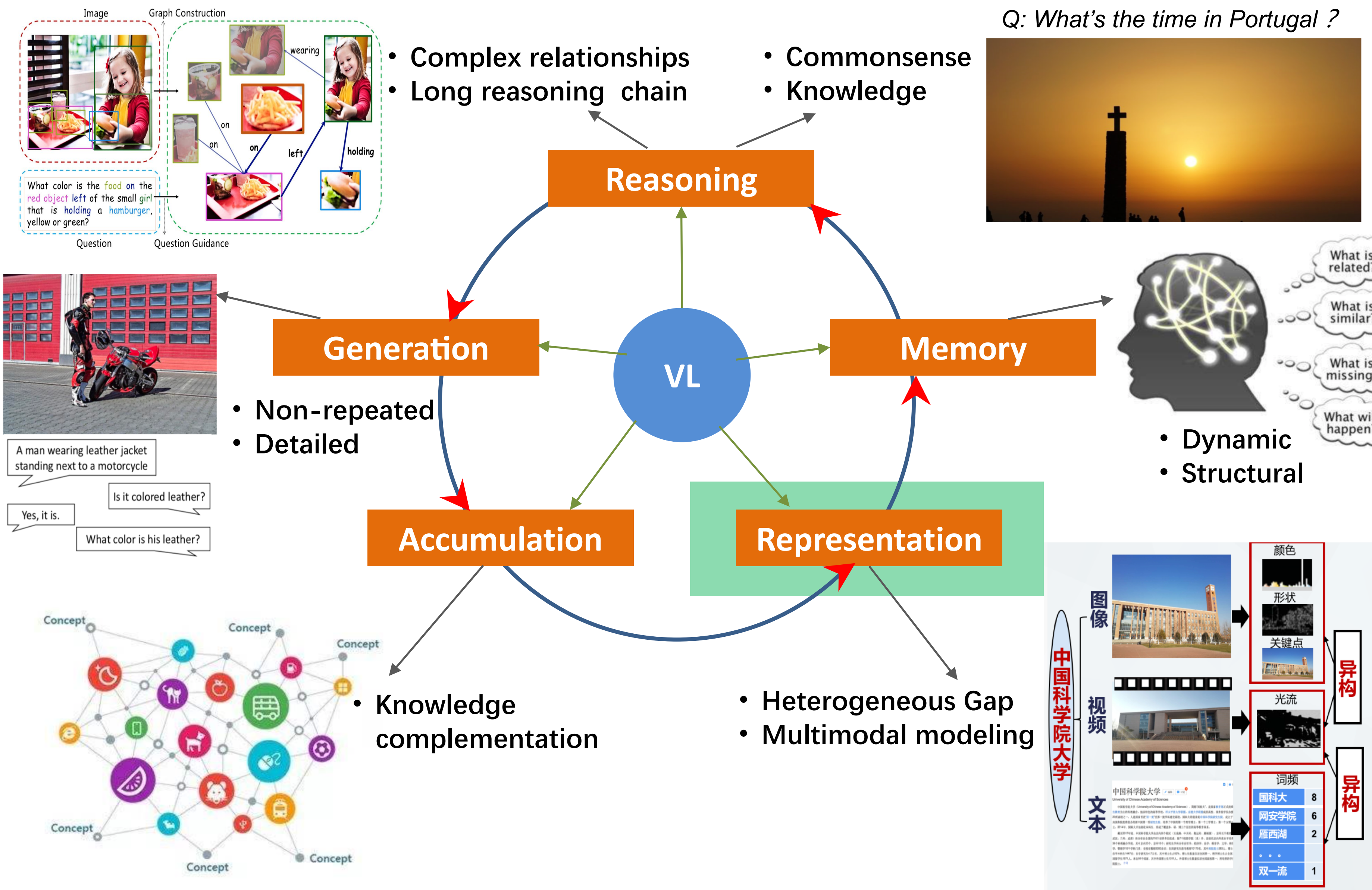


Image credit: https://mp.weixin.qq.com/s/HKjMqZ6bxnUoZg2WlBapJg

Our Works Augmented by Prior Human Knowledge



Our Works Augmented by Prior Human Knowledge

VQA/ Visual Dialogue/Scene Graph Generation 2020

- **Jing Yu**, Xiaoze Jiang, Yue Hu, Qi Wu, et al. *Learning Dual Encoding Model for Adaptive Visual Understanding in Visual Dialogue*, **TIP 2020**
- **Jing Yu**, Zihao Zhu, Yujing Wang, Yue Hu, et al. *Cross-modal knowledge reasoning for knowledge-based visual question answering*, **Pattern Recognition 2020**
- **Jing Yu**, Weifeng Zhang, Yuhang Lu, Qi Wu, et al. *Reasoning on the Relation: Enhancing Visual Representation for Visual Question Answering and Cross-modal Retrieval*, **TMM 2020**
- **Jing Yu**, Weifeng Zhang, Zhuoqian Yang, Yue Hu, Qi Wu. *Cross-modal learning with prior visual relation knowledge*, **Knowledge-based Systems 2020**
- **Jing Yu**, Yuan Chai, Yue Hu, Qi Wu. *CogTree: Cognition Tree Loss for Unbiased Scene Graph Generation*. https://arxiv.org/abs/2009.07526
- Weifeng Zhang, **Jing Yu**, Hua Hu, Haiyang Hu. *Multimodal feature fusion by relational reasoning and attention for visual question answering*, **Information Fusion 2020**
- Xiaoze Jiang, **Jing Yu***, Yue Hu, Qi Wu, et al. *Deep Visual Understanding Like Humans: An Adaptive Dual Encoding Model for Visual Dialogue*, **AAAI 2020**
- Zihao Zhu*, **Jing Yu*(Equal)**, Yujing Wang, Yue Hu, Qi Wu, et al. *Mucko: Multi-Layer Cross-Modal Knowledge Reasoning for Fact-based Visual Question Answering*, **IJCAI 2020**
- Xiaoze Jiang*, **Jing Yu*(Equal)**, Xingxing Zhang, Yue Hu, Qi Wu, et al. *DAM: Deliberation, Abandon and Memory Networks for Generating Detailed and Non-repetitive Responses in Visual Dialogue*, **IJCAI 2020**

Our Works Augmented by Prior Human Knowledge

- Cognition-Guided Scene Graph Generation (SGG)
  - Cognition-based induction for unbiased SGG

- Cognition-Guided Visual and Non-Visual Representation
  - Visual understanding with relational visual and non-visual semantics
  - Visual understanding from multiple views and grains

CogTree: Cognition Tree Loss for Unbiased Scene Graph Generation

https://arxiv.org/abs/2009.07526

Jing Yu*, Yuan Chai, Yue Hu, Qi Wu

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

中国科学院大学
University of Chinese Academy of Sciences

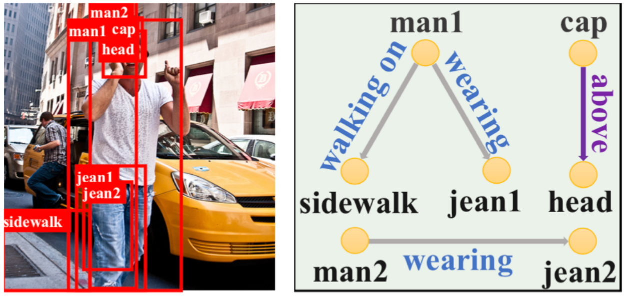
北京航空航天大學
BEIHANG UNIVERSITY

THE UNIVERSITY of ADELAIDE

Scene Graph Generation

Heterogeneous Gap between Vision and Language

Structural Representation → • Scene Graph Generation (SGG)

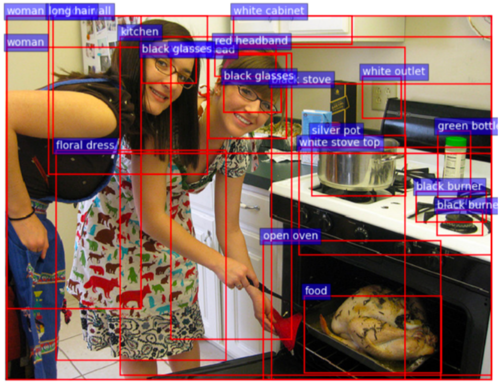


↑

Relationship Understanding → • Visual Relationship Detection (VRD)

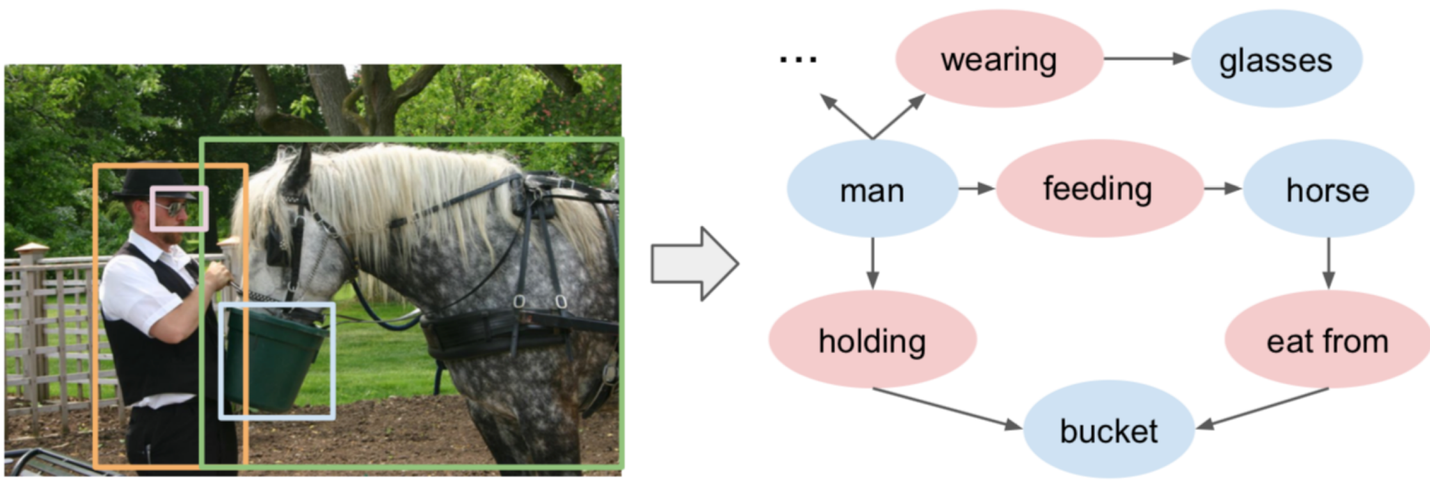
↑

Object Understanding →
- Object Detection
- Object Recognition
- Attribute Recognition

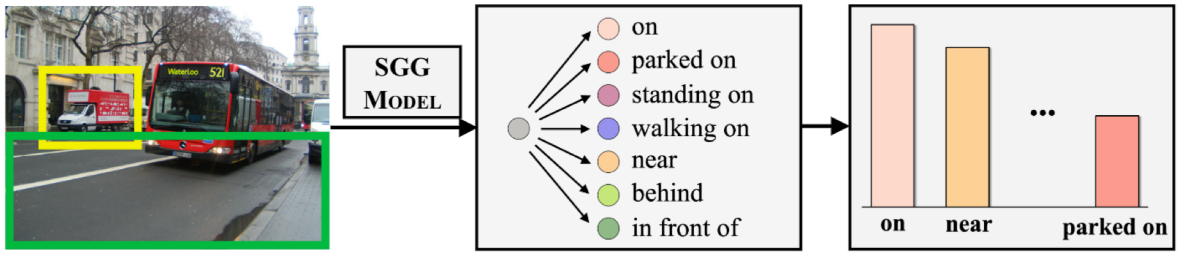


Unbiased Scene Graph Generation

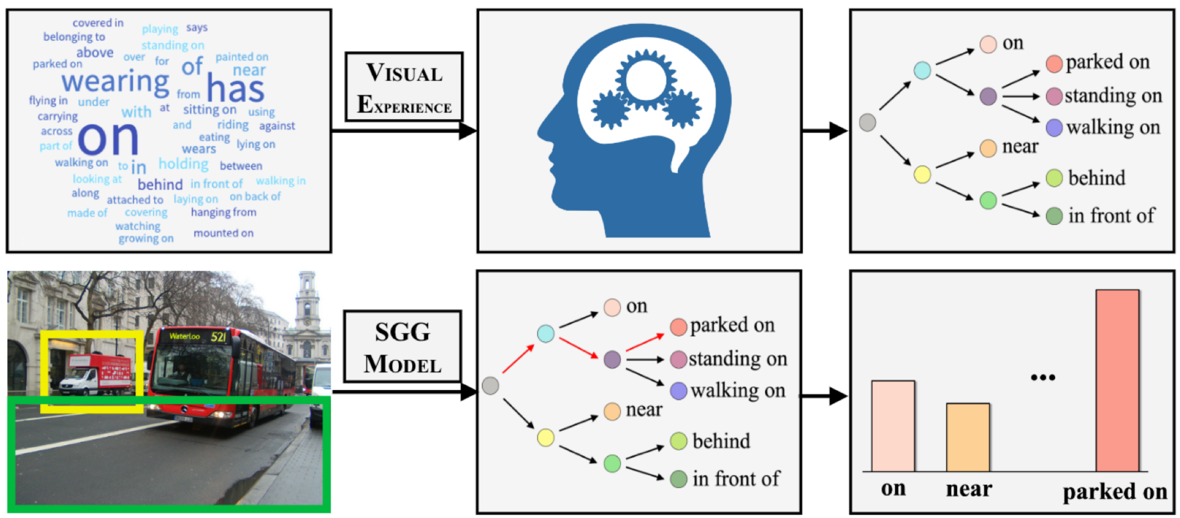
- Most methods generates biased scene graphs
  - Predominantly generate "head" relationships
  - Lack of fine-grained semantics



Think from cognition view



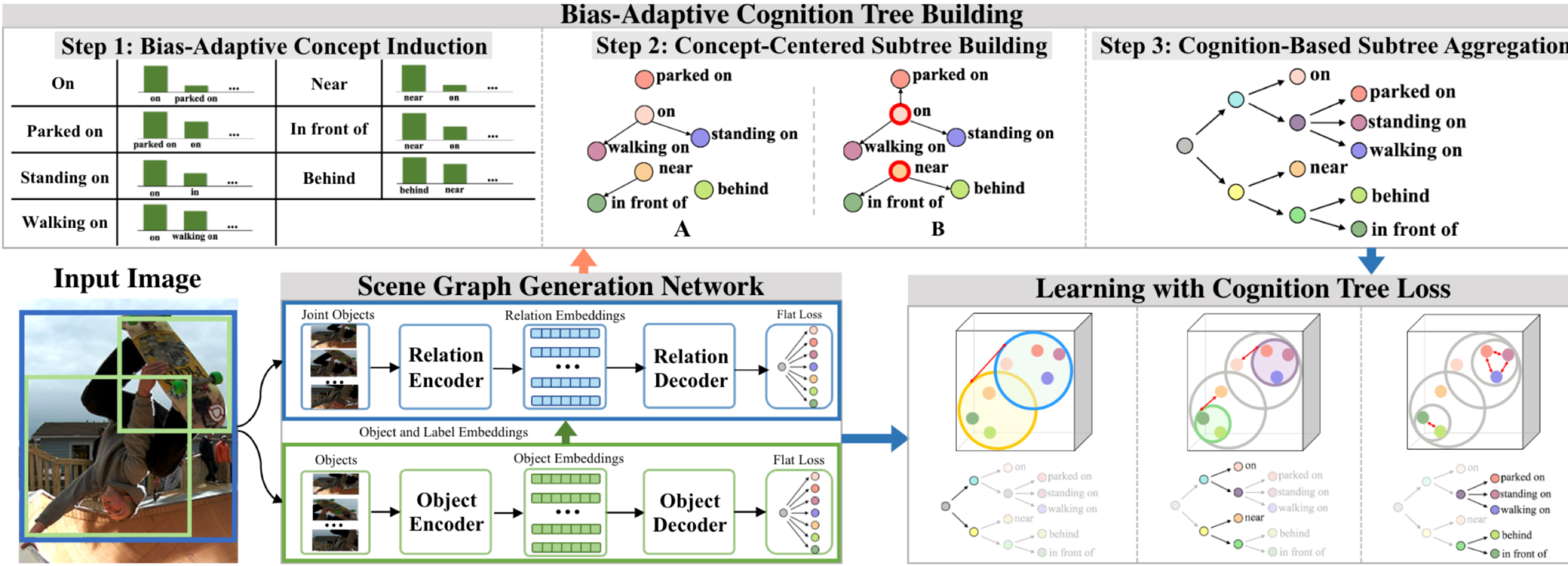
(a) Flat thinking.



(b) Cognition-based hierarchical thinking.

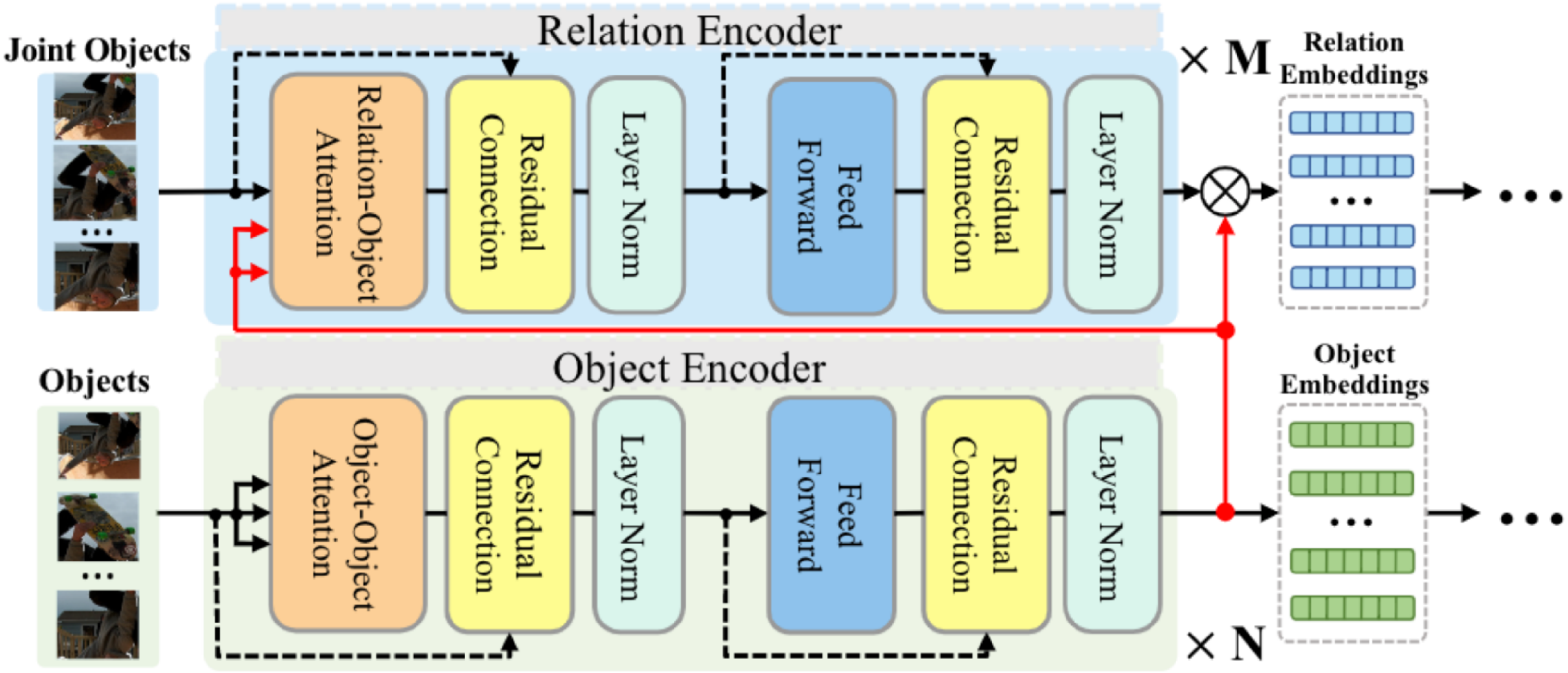
Proposed CogTree Loss

- SGG Network ➔ Bias-Adaptive Cognition Tree Building ➔ Learning with Cognition Tree Loss



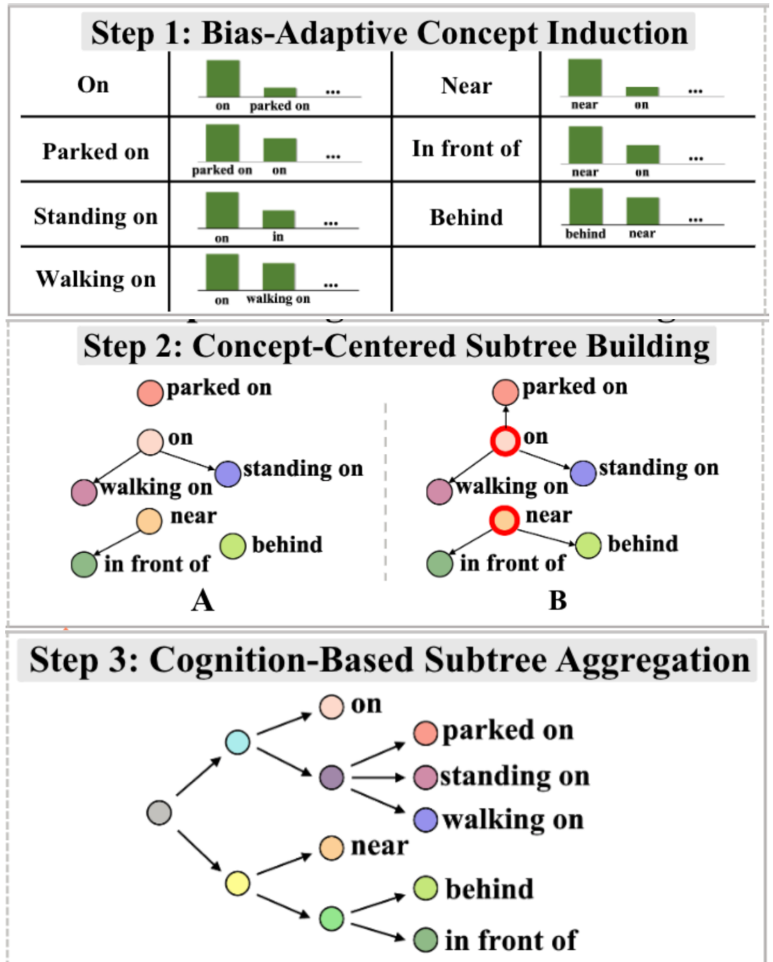
Proposed CogTree Loss

- SGG Network ➔ Bias-Adaptive Cognition Tree Building ➔ Learning with Cognition Tree Loss



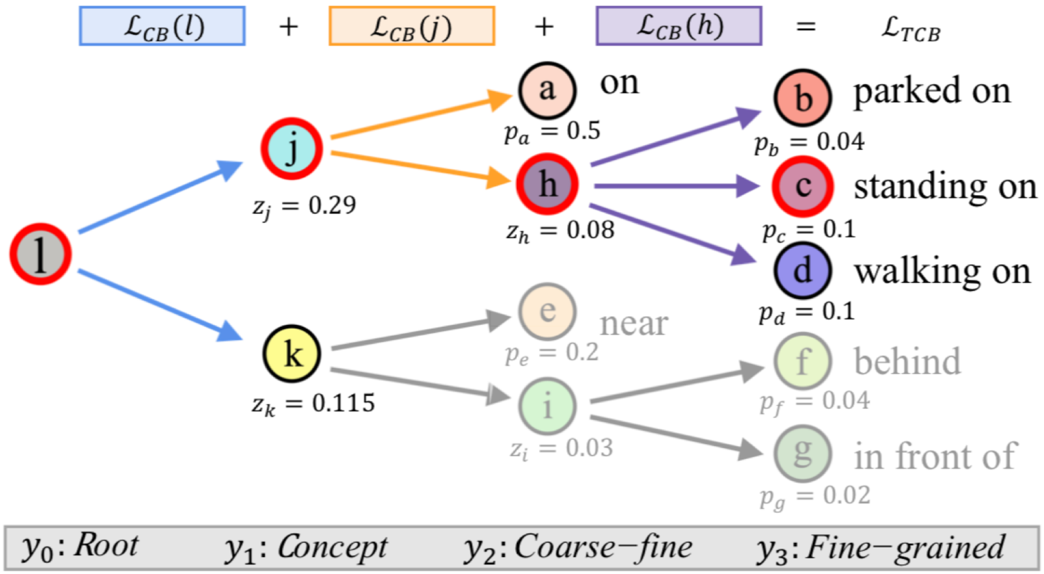
Proposed CogTree Loss

- SGG Network ➔ Bias-Adaptive Cognition Tree Building ➔ Learning with Cognition Tree Loss



Proposed CogTree Loss

- SGG Network ➔ Bias-Adaptive Cognition Tree Building ➔ Learning with Cognition Tree Loss



$$\mathcal{L}_{TCB} = \frac{1}{|L_{path}|} \sum_{i \in L_{path}} -w_i \log\left(\frac{\exp(z_i)}{\sum_{z_j \in B(i)} \exp(z_j)}\right)$$

$$\mathcal{L}_{CB} = -w_k \log\left(\frac{\exp(p_k)}{\sum_{p_j \in P_{pred}} \exp(p_j)}\right)$$

$$\mathcal{L} = \mathcal{L}_{CB} + \lambda \mathcal{L}_{TCB}$$

Experiments

- Comparison with state-of-the-art models
  - Baselines
    - MOTIFS, VCTree, SG-transformer, IMP+, FREO, KERN
    - Debiasing approaches: Focal, Reweighting, Resampling, TDE
  - Our X + CogTree > X+debiased approaches > exsiting baselines

	Scene Graph Detection			Scene Graph Classification			Predicate Classification		
Model	mR@20	mR@50	mR@100	mR@20	mR@50	mR@100	mR@20	mR@50	mR@100
IMP+	-	3.8	4.8	-	5.8	6.0	-	9.8	10.5
FREQ	4.5	6.1	7.1	5.1	7.2	8.5	8.3	13.0	16.0
KERN	-	6.4	7.3	-	9.4	10.0	-	17.7	19.2
MOTIFS	4.2	5.7	6.6	6.3	7.7	8.2	10.8	14.0	15.3
VCTree	5.2	6.9	8.0	8.2	10.1	10.8	14.0	17.9	19.4
MOTIFS (baseline)	4.1	5.5	6.8	6.5	8.0	8.5	11.5	14.6	15.8
MOTIFS + Focal	3.9	5.3	6.6	6.3	8.0	8.5	11.5	14.6	15.8
MOTIFS + Reweight	6.5	8.4	9.8	8.4	10.1	10.9	16.0	20.0	21.9
MOTIFS + Resample	5.9	8.2	9.7	9.1	11.0	11.8	14.7	18.5	20.0
MOTIFS + TDE	5.8	8.2	9.8	9.8	13.1	14.9	18.5	25.5	**29.1**
MOTIFS + CogTree	**7.9**	**10.4**	**11.8**	**12.1**	**14.9**	**16.1**	**20.9**	**26.4**	29.0
VCTree (baseline)	4.2	5.7	6.9	6.2	7.5	7.9	11.7	14.9	16.1
VCTree + TDE	6.9	9.3	11.1	8.9	12.2	14.0	18.4	25.4	28.7
VCTree + CogTree	**7.8**	**10.4**	**12.1**	**15.4**	**18.8**	**19.9**	**22.0**	**27.6**	**29.7**
SG-transformer (baseline)	5.6	7.7	9.0	8.6	11.5	12.3	14.4	18.5	20.2
SG-transformer + CogTree	**7.9**	**11.1**	**12.7**	**13.0**	**15.7**	**16.7**	**22.9**	**28.4**	**31.0**

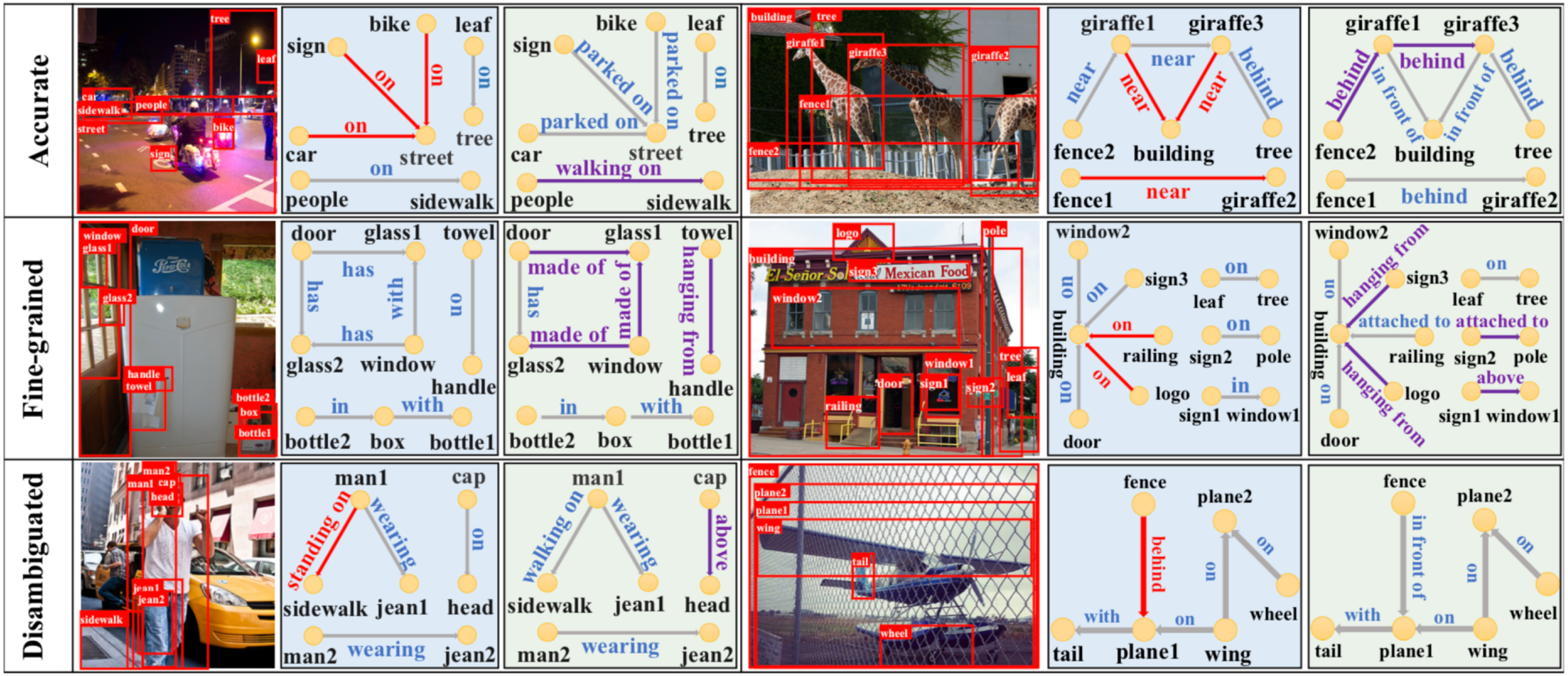
Experiments

- Ablation Study
  - Both CogTree and balancing losses benefit the performance
  - Tree Structure matters
  - Weighting strategy matters

		Scene Graph Detection			Scene Graph Classification			Predicate Classification		
Method		mR@20	mR@50	mR@100	mR@20	mR@50	mR@100	mR@20	mR@50	mR@100
CogTree + $\mathcal{L}$ (full model)		**7.92**	**11.05**	**12.70**	**12.96**	**15.68**	**16.72**	**22.89**	**28.38**	**30.97**
1	CogTree + $\mathcal{L}_{TCB}$	7.70	10.39	12.07	12.15	15.07	16.15	21.08	27.08	29.41
2	CogTree + $\mathcal{L}_{TCE}$	7.57	10.53	11.86	12.14	14.42	15.29	21.16	26.14	28.32
3	$\mathcal{L}_{CB}$	6.74	9.56	11.29	10.76	13.13	13.88	18.02	23.40	25.25
4	$\mathcal{L}_{CE}$	5.55	7.74	8.98	8.57	11.46	12.27	14.35	18.48	20.21
5	Fuse-layer + $\mathcal{L}$	5.86	8.02	9.05	8.17	10.39	11.32	13.77	18.87	20.77
6	Fuse-subtree + $\mathcal{L}$	5.36	7.19	8.28	8.71	10.66	11.61	16.20	20.17	22.12
7	Cluster-tree + $\mathcal{L}$	5.84	8.10	9.12	8.86	10.88	11.52	15.12	19.20	20.81
8	CogTree + $\mathcal{L}$(MAX)	5.38	7.16	8.16	8.97	10.85	11.83	15.48	19.93	21.87
9	CogTree + $\mathcal{L}$(SUM)	1.86	3.09	3.68	6.58	8.82	9.86	11.31	15.67	17.98

Experiments

- Visualization
  - CogTree predicts more fine-grained relationships.
  - CogTree successfully distinguishes visually and semantically similar relationships.



Mucko: Multi-Layer Cross-Modal Knowledge Reasoning for Fact-based Visual Question Answering

IJCAI 2020

https://github.com/astro-zihao/mucko.

Zihao Zhu, **Jing Yu***, Yujing Wang, Yajing Sun, Yue Hu, Qi Wu

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

中国科学院大学
University of Chinese Academy of Sciences

Microsoft Research
微软亚洲研究院

THE UNIVERSITY of ADELAIDE

KB-VQA

- Previous Visual Question Answering tasks only need perception
  - Language or visual bias
  - Short reasoning chain (salient objects, simple relationships, few context, limited attributes)
  - Few commonsense & knowledge (pure visual)

➡ FVQA (2018, TPAMI, Univ. of Adelaide)

Question:
What is the red cylinder object in the image is used for?

Factual Knowledge:
<fire hydrant, UsedFor, firefighting>

➡ OK-VQA (2019, CVPR, CMU)

Q: Which American president is associated with the stuffed animal seen here?

A: Teddy Roosevelt

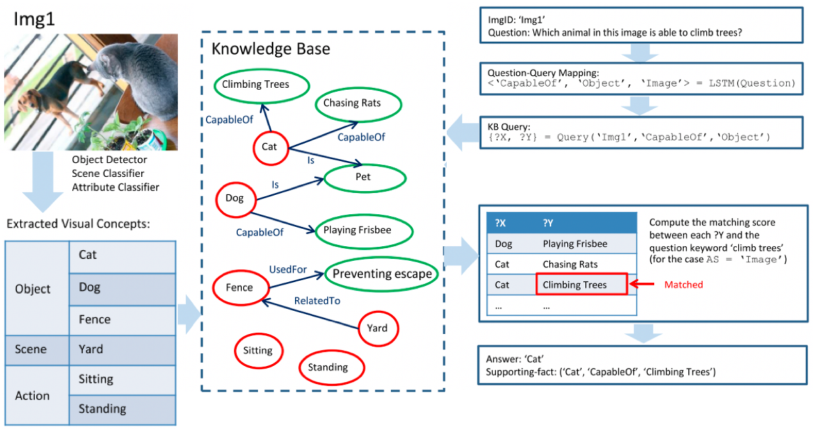
Outside Knowledge

Another lasting, popular legacy of Roosevelt is the stuffed toy bears—teddy bears—named after him following an incident on a hunting trip in Mississippi in 1902.

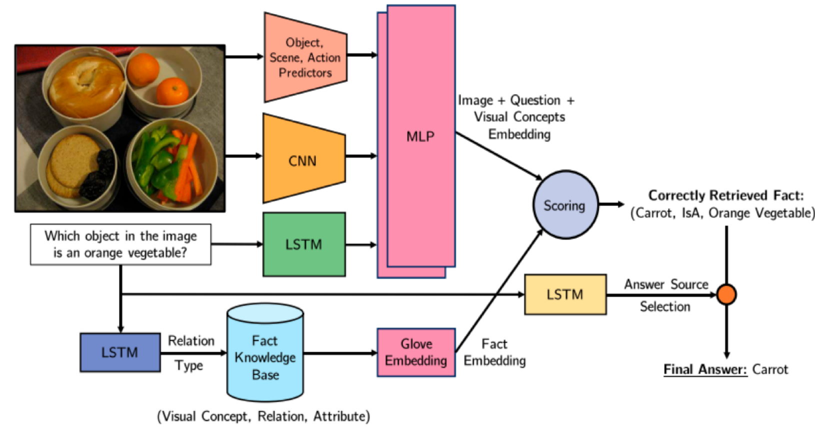
Developed apparently simultaneously by toymakers ... and named after President Theodore "Teddy" Roosevelt, the teddy bear became an iconic children's toy, celebrated in story, song, and film.

At the same time in the USA, Morris Michtom created the first teddy bear, after being inspired by a drawing of Theodore "Teddy" Roosevelt with a bear cub.

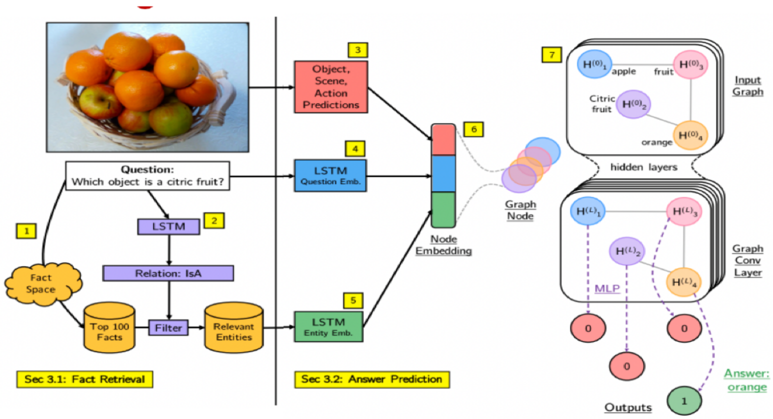
Related Work



▲ Retrieval-based Approach [1]



▲ Fusion-based Approach [2]



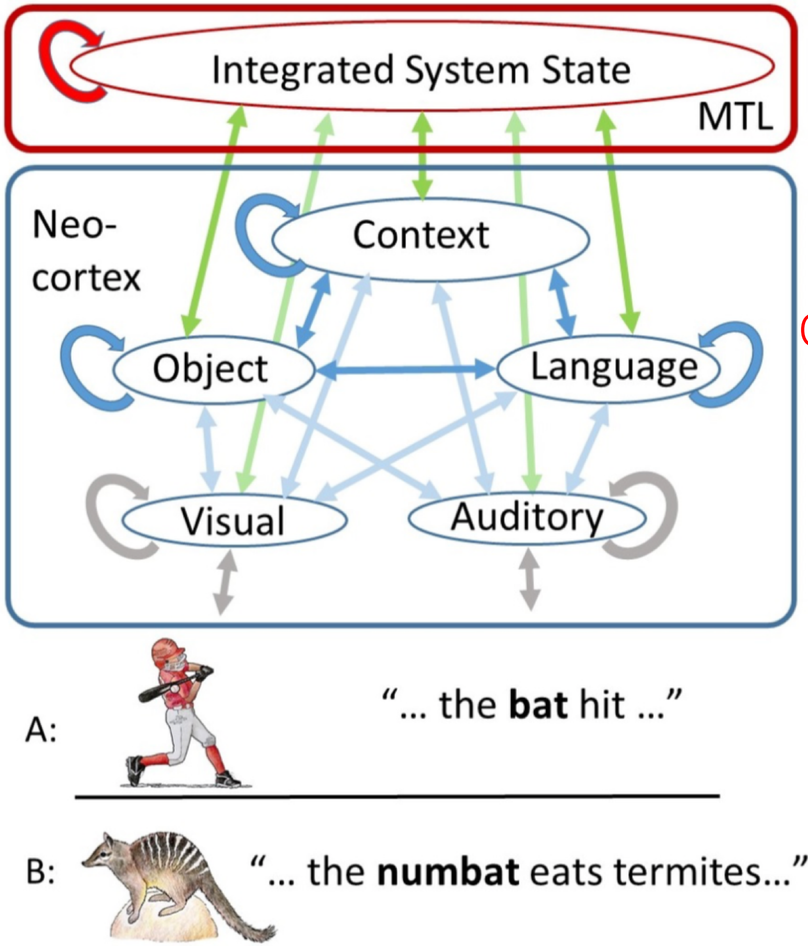
▲ Global Evaluation [3]

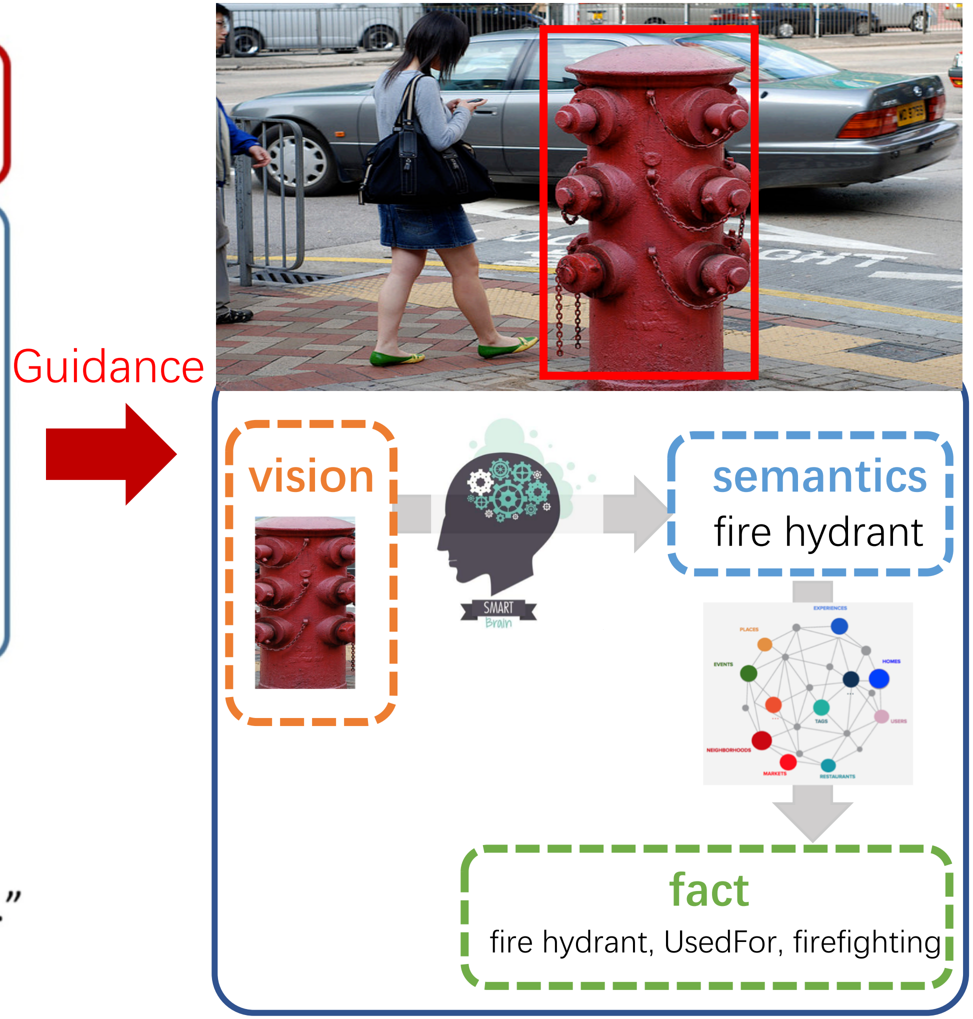
The complementary role of visual-semantic-knowledge has been ignored！

[1]Peng Wang et al. TPAMI 2018 FVQA: Fact-based Visual Question Answering
[2]Narasimhan et al. ECCV 2018 Straight to the Facts: Learning Knowledge Base Retrieval for Factual Visual Question Answering
[3] Narasimhan et al. NIPS 2018 Out of the Box: Reasoning with Graph Convolution Nets for Factual Visual Question Answering

Think from cognition view

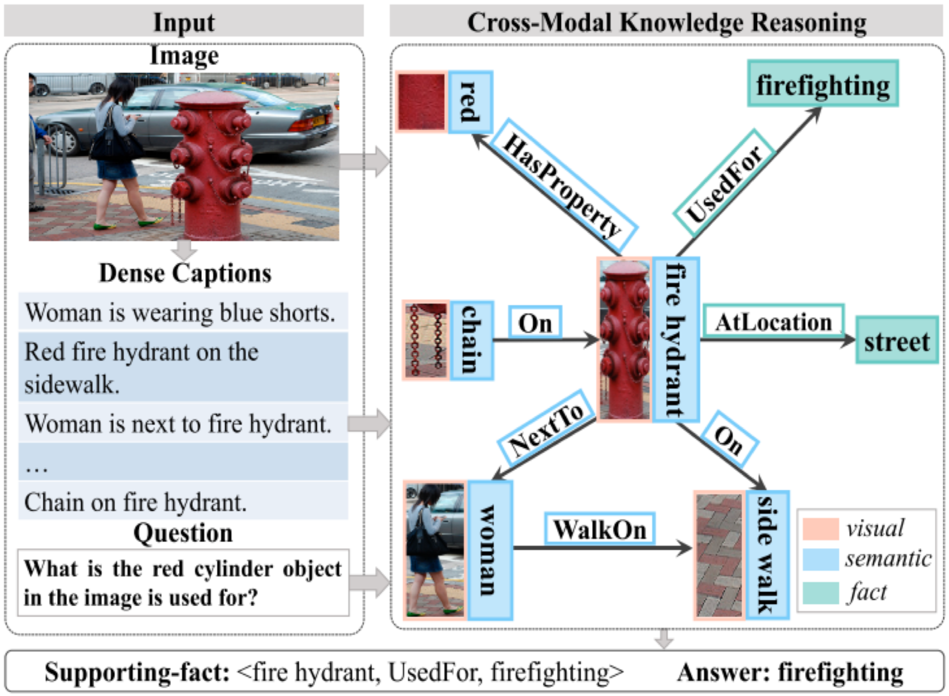
brain's understanding system





[James L. McClelland et al. arxiv 2020 Extending Machine Language Models toward Human-Level Language Understanding]

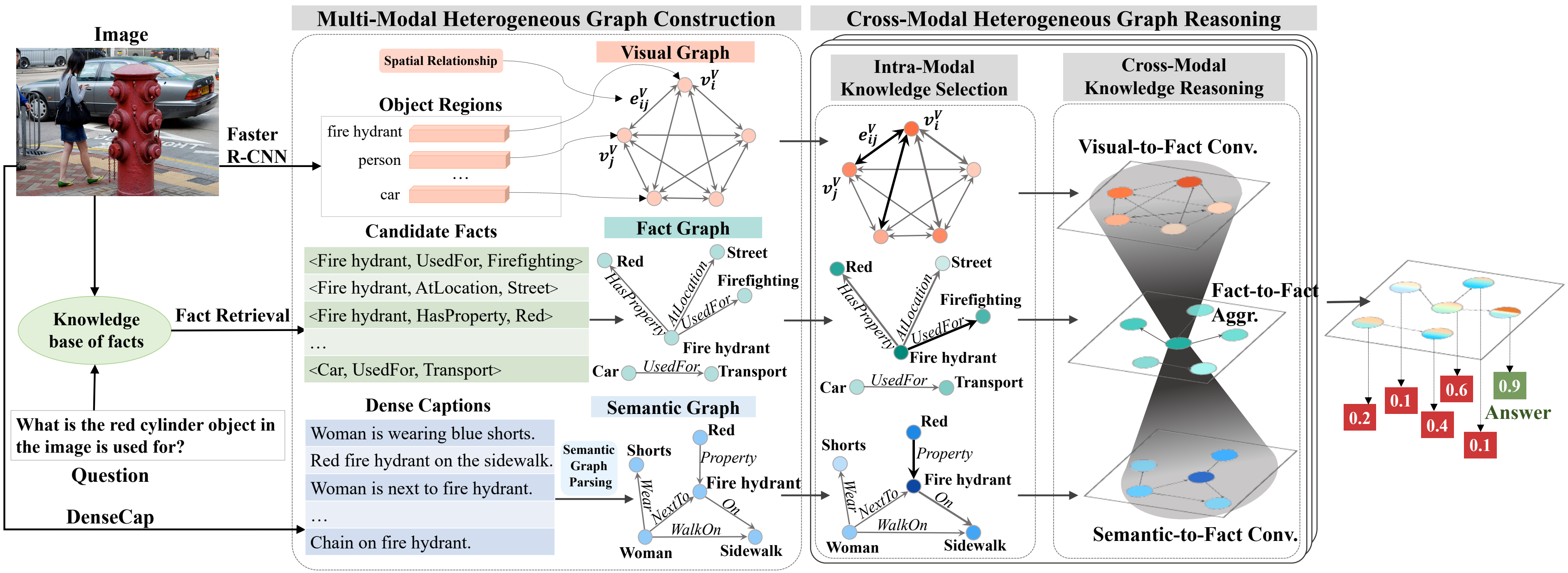
Multi-Layer Cross-Modal Knowledge Reasoning (Mucko)



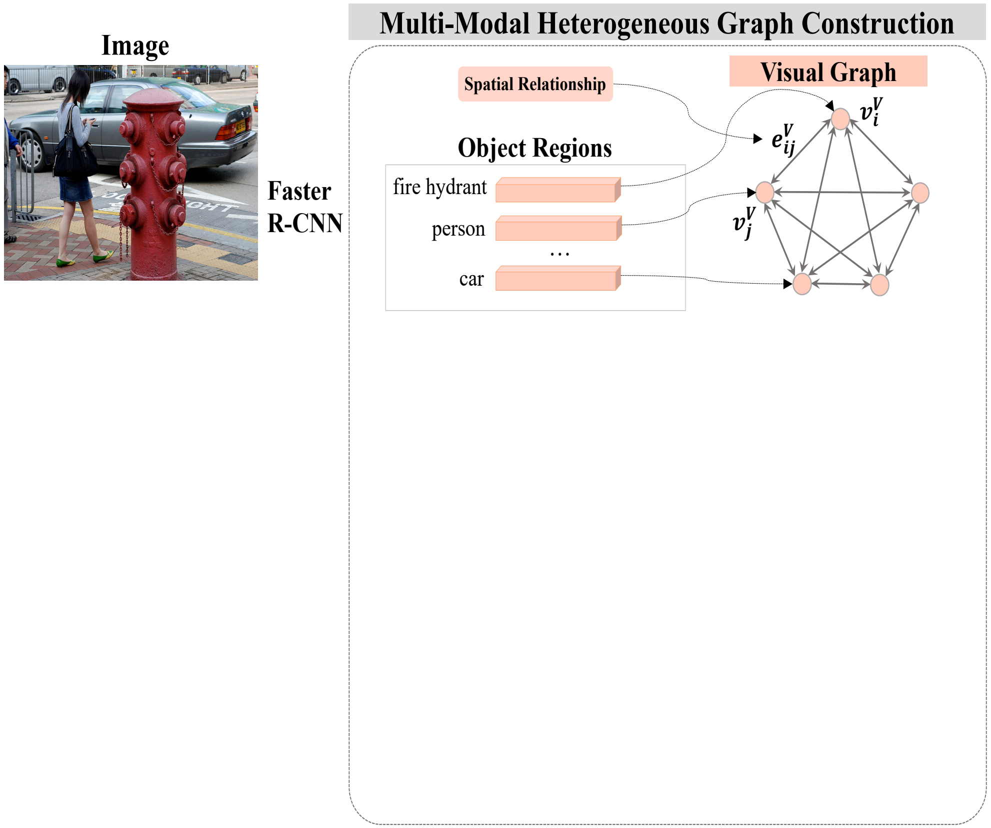
- **Multi-layer graph representation**
  - **visual layer:** object appearance and visual relationships
  - **semantic layer:** high-level abstraction
  - **fact layer:** knowledge of facts
- **Heterogeneous graph convolutional network**
  adaptively collect complementary evidence in the multi-layer graphs.

Multi-Layer Cross-Modal Knowledge Reasoning (Mucko)

- Multi-Modal Heterogenous Graph Construction ➔ Intra-Modal Knowledge Selection➔ Cross-Modal Knowledge Reasoning



- Multi-Modal Heterogenous Graph Construction ➔ Intra-Modal Knowledge Selection➔ Cross-Modal Knowledge Reasoning

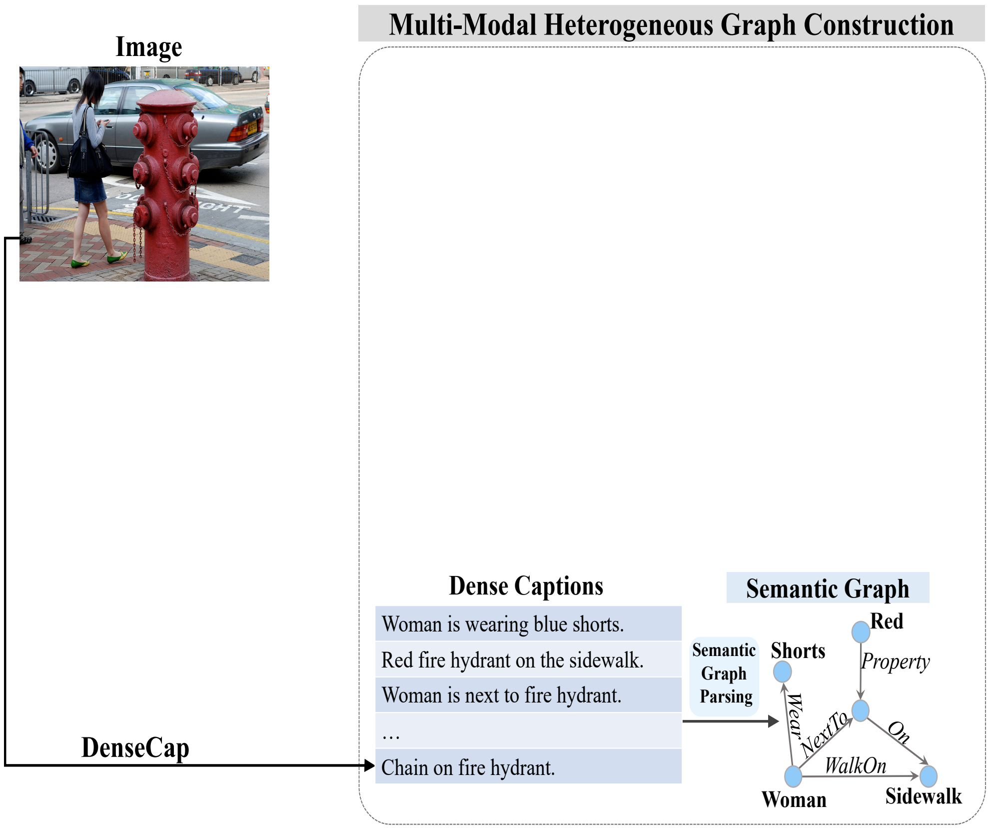


Visual Graph

- Faster-RCNN is used to extract a set of objects, $O = \{o_i\}_{i=1}^{K}(K = 36)$.
- Each object has a 2048-d feature.
- Construct a visual graph $G^V = (V^V, E^V)$ over O
- $v_i^V \in R^{2048}$
- spatial relationship $r_i^V = \left[\frac{x_j - x_i}{w_i}, \frac{y_j - y_i}{h_i}, \frac{w_j}{w_i}, \frac{h_j}{h_i}, \frac{w_j h_j}{w_i h_i}\right]$

Multi-Layer Cross-Modal Knowledge Reasoning (Mucko)

- Multi-Modal Heterogenous Graph Construction ➔ Intra-Modal Knowledge Selection➔ Cross-Modal Knowledge Reasoning

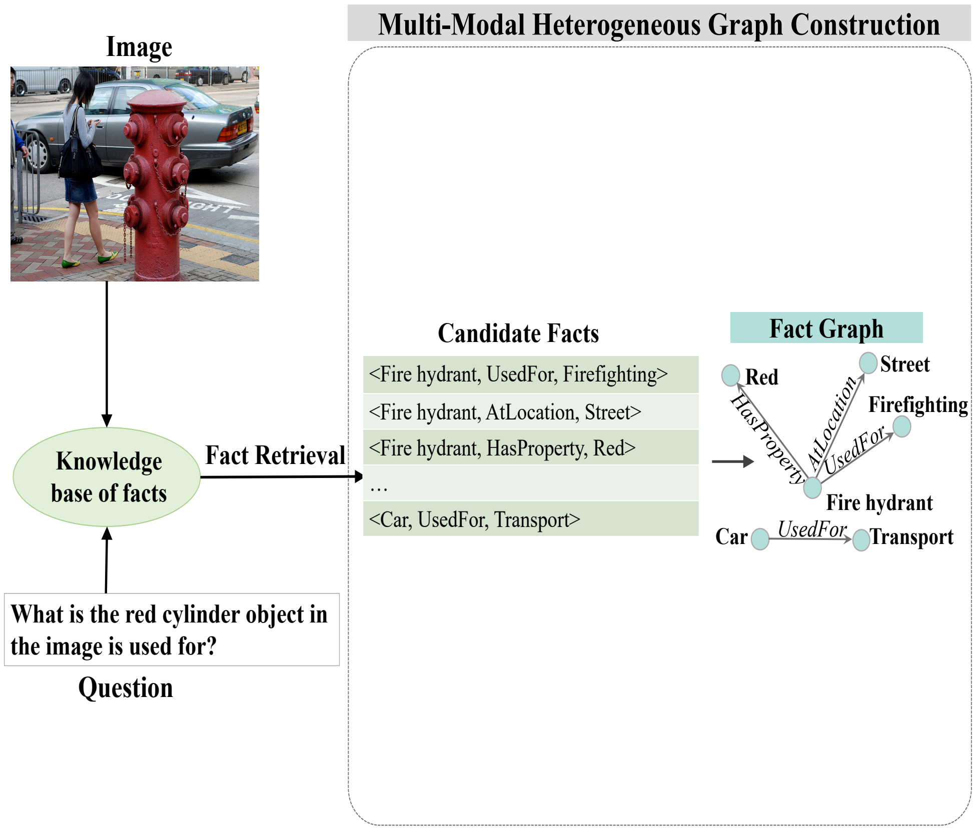


Semantic Graph

- **DenseCap** is used to generate dense captions about image.
- **SPICE** is used to convert text into semantic graph $G^S = (V^S, E^S)$.
- Each node and edge is represented by **GloVe** embedding.
- $v_i^s \in R^{300}$
- $r_i^s \in R^{300}$

Multi-Layer Cross-Modal Knowledge Reasoning (Mucko)

- Multi-Modal Heterogenous Graph Construction ➔ Intra-Modal Knowledge Selection ➔ Cross-Modal Knowledge Reasoning

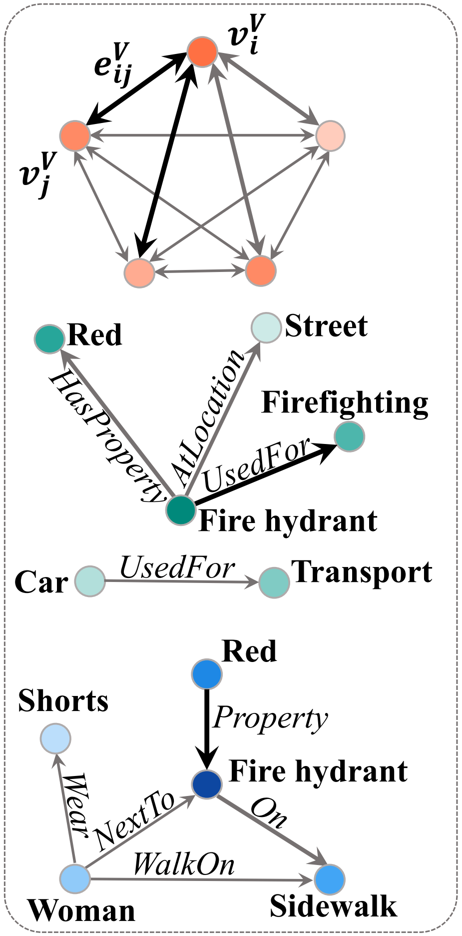


Fact Graph

- for each *fact <e1, r, e2>* of KB, compute the cosine similarities of (e1,e2) and (o1,o2,⋯o36)
- average these similarities to assign a score to the *fact*
- sort and select top-k facts according to scores.
- train a relation classifier to predict relation type based on the question
- filter the facts according to relation type.

Multi-Layer Cross-Modal Knowledge Reasoning (Mucko)

- Multi-Modal Heterogenous Graph Construction ➔ Intra-Modal Knowledge Selection ➔ Cross-Modal Knowledge Reasoning



Question-guided Node Attention: evaluate the relevance of each node corresponding to the question by at- tention mechanism.

$$\alpha_i = \text{softmax}(\boldsymbol{w}_a^T \tanh(\mathbf{W}_1 \boldsymbol{v}_i + \mathbf{W}_2 \boldsymbol{q}))$$

Question-guided Edge Attention: evaluate the importance of edge constrained by the neighbor node v_j regarding to v_i as:

$$\beta_{ji} = \text{softmax}(\boldsymbol{w}_b^T \tanh(\mathbf{W}_3 \boldsymbol{v}_j' + \mathbf{W}_4 \boldsymbol{q}'))$$

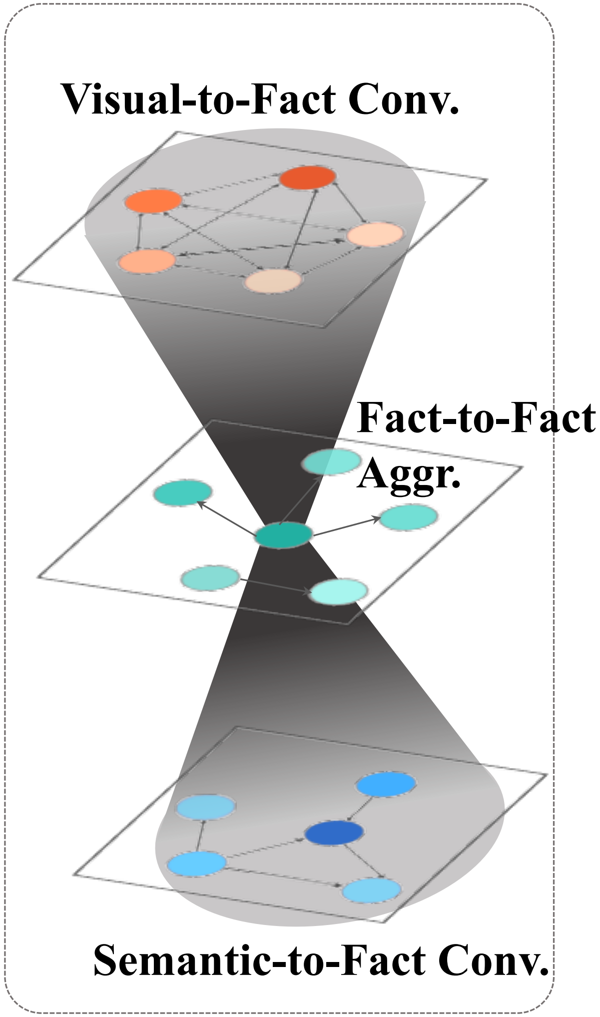
Intra-Modal Graph Convolution: gather the neighborhood information and update the representation of vi as:

$$\boldsymbol{m}_i = \sum_{j \in \mathcal{N}_i} \beta_{ji} \boldsymbol{v}_j'$$

$$\hat{\boldsymbol{v}}_i = \text{ReLU}(\mathbf{W}_7 [\boldsymbol{m}_i, \alpha_i \boldsymbol{v}_i])$$

Multi-Layer Cross-Modal Knowledge Reasoning (Mucko)

- Multi-Modal Heterogenous Graph Construction ➔ Intra-Modal Knowledge Selection➔ Cross-Modal Knowledge Reasoning



Visual-to-Fact Convolution: gather complementary information from visual graph by cross-modal convolutions.

$$\gamma_{ji}^{V\text{-}F} = \mathrm{softmax}(\boldsymbol{w}_c \tanh(\mathbf{W}_8 \hat{\boldsymbol{v}}_j^V + \mathbf{W}_9[\hat{\boldsymbol{v}}_i^F, \boldsymbol{q}]))$$

Semantic-to-Fact Convolution: gather complementary information from semantic graph by cross-modal convolutions.

Fact-to-Fact Aggregation: gather information from the fact graph by intra-modal convolutions.

Experiments——SOTA Comparison

State-of-the-art comparison on FVQA

Method	Overall Accuracy	
	top-1	top-3
LSTM-Question+Image+Pre-VQA [10]	24.98	40.40
Hie-Question+Image+Pre-VQA [10]	43.14	59.44
FVQA (top-3-QQmaping) [10]	56.91	64.65
FVQA (Ensemble) [10]	58.76	-
Straight to the Facts (STTF) [9]	62.20	75.60
Reading Comprehension [6]	62.96	70.08
Out of the Box (OB) [8]	69.35	80.25
Human [10]	77.99	-
Mucko	**73.06** ± 0.39	**85.94** ± 0.46

top-1: ↑ **3.7%**
top-3: ↑ **5.7%**

State-of-the-art comparison on Visual7w+KB

Method	Overall Accuracy	
	top-1	top-3
KDMN-NoKnowledge [5]	45.1	-
KDMN-NoMemory [5]	51.9	-
KDMN [5]	57.9	-
KDMN-Ensemble [5]	60.9	-
Out of the Box (OB) [1] [8]	57.32	71.61
Mucko (ours)	**68.88** ± 0.52	**85.13** ± 0.67

top-1: ↑ **11.5%**
top-3: ↑ **13.5%**

State-of-the-art comparison on OK-VQA

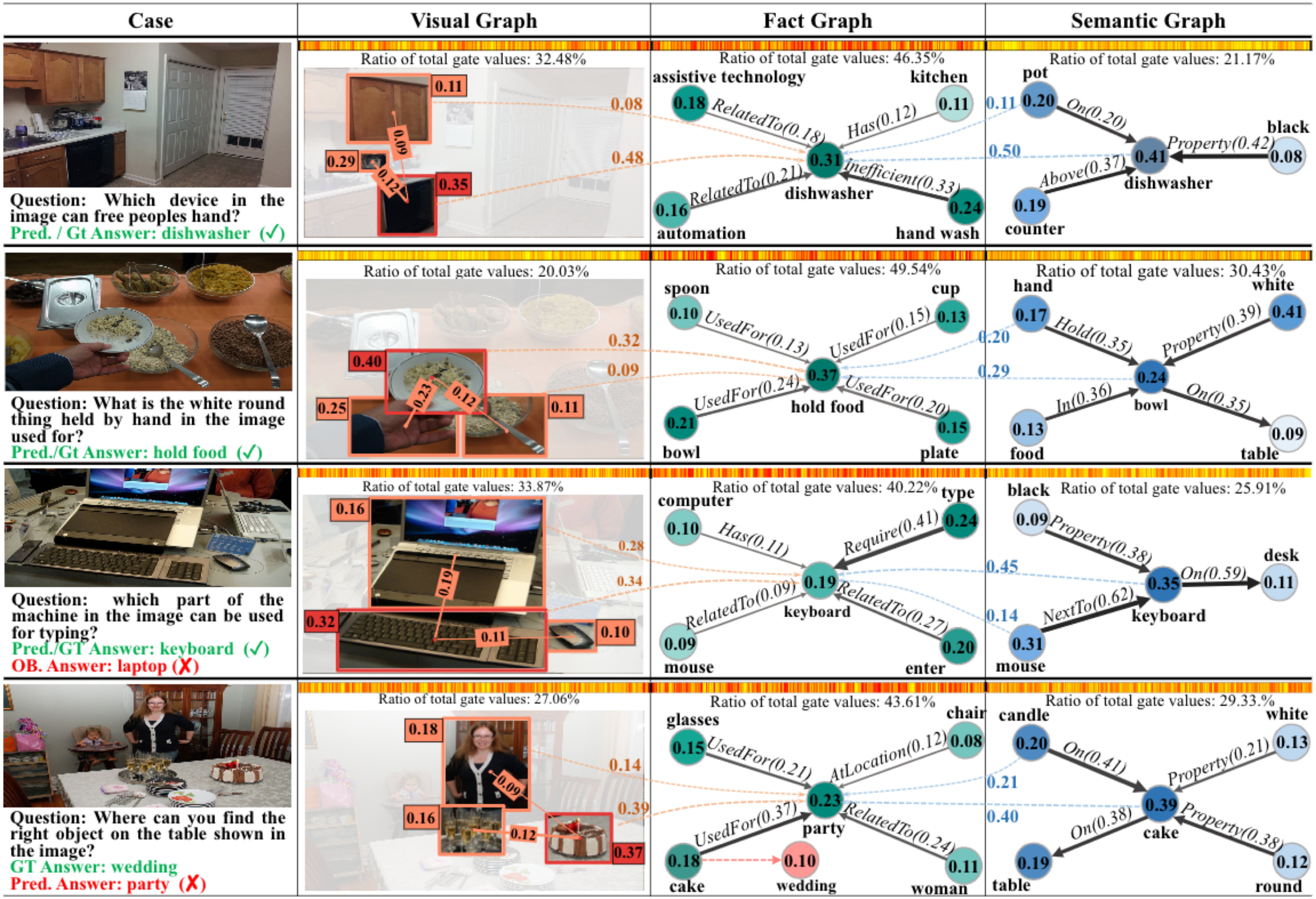
Method	Overall Accuracy	
	top-1	top-3
Q-Only [7]	14.93	-
MLP [7]	20.67	-
BAN [3]	25.17	-
MUTAN [1]	26.41	-
ArticleNet (AN) [7]	5.28	-
BAN + AN [7]	25.61	-
MUTAN + AN [7]	27.84	-
BAN/AN oracle [7]	27.59	-
MUTAN/AN oracle [7]	28.47	-
Mucko (ours)	**29.20** ± 0.31	**30.66** ± 0.55

top-1: ↑ **0.7%**

Experiments——Ablation Study

Method		Overall Accuracy	
		top-1	top-3
Mucko (full model)		**73.06**	**85.94**
1	w/o Intra-Modal Knowledge Selection	70.50	81.77
2	w/o Semantic Graph	71.28	82.76
3	w/o Visual Graph	69.12	78.05
4	w/o Semantic Graph & Visual Graph	20.43	29.10
5	S-to-F Concat.	67.82	76.65
6	V-to-F Concat.	69.93	80.12
7	V-to-F Concat. & S-to-F Concat.	70.68	82.04
8	w/o relationships	72.10	83.75

Experiments——Qualitative Analysis



Learning Dual Encoding Model for Adaptive Visual Understanding in Visual Dialogue

TIP 2020

https://github.com/JXZe/DualVD

Jing Yu*, Xiaoze Jiang, Zengchang Qin, Weifeng Zhang, Yue Hu, Qi Wu

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

中国科学院大学
University of Chinese Academy of Sciences

北京航空航天大學
BEIHANG UNIVERSITY

THE UNIVERSITY of ADELAIDE

Visual Dialogue

Captioning

Two people are in a wheelchair and one is holding a racket.

VQA

Q: How many people on wheelchairs ?
A: Two

Q: How many wheelchairs ?
A: One

Visual Dialog

Q: How many people are on wheelchairs ?
A: Two
Q: What are their genders ?
A: One male and one female
Q: Which one is holding a racket ?
A: The woman

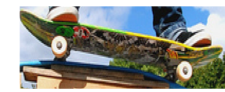
Motivation

Image

C: A man doing a grind on a skateboard.
Q1: Is the man on the skateboard?
A1: Yes, he is.
…
Q4: Is he younger or older?
A4: He is in the middle-aged.
Q5: Is there sky in the picture?
A5: Yes, the sky is deep blue with some clouds.

History

the man



skateboard

Prospect

- Visual Dialogue task demands the agent to adaptively focus on diverse visual content with respect to the current question.
- The **key challenge** in Visual Dialogue task is thus to learn a more comprehensive and semantic-rich image representation, which may have adaptive attentions on the image for variant questions.

Motivation

Image

C: A man doing a grind on a skateboard.
Q1: Is the man on the skateboard?
A1: Yes, he is.
…
Q4: Is he younger or older?
A4: He is in the middle-aged.
Q5: Is there sky in the picture?
A5: Yes, the sky is deep blue with some clouds.

History

sky

Background

- Visual Dialogue task demands the agent to adaptively focus on diverse visual content with respect to the current question.
- The **key challenge** in Visual Dialogue task is thus to learn a more comprehensive and semantic-rich image representation, which may have adaptive attentions on the image for variant questions.

Motivation

Image

C: A man doing a grind on a skateboard.
Q1: Is the man on the skateboard?
A1: Yes, he is.
...
Q4: Is he younger or older?
A4: He is in the middle-aged.
Q5: Is there sky in the picture?
A5: Yes, the sky is deep blue with some clouds.

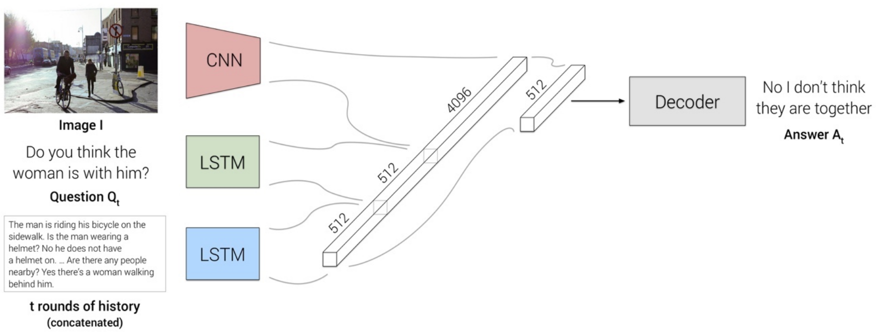
History

the middle-aged man

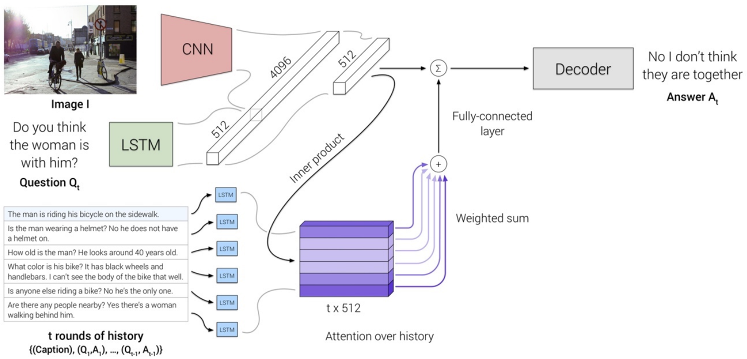
Higher-level semantics

- Visual Dialogue task demands the agent to adaptively focus on diverse visual content with respect to the current question.
- The **key challenge** in Visual Dialogue task is thus to learn a more comprehensive and semantic-rich image representation, which may have adaptive attentions on the image for variant questions.

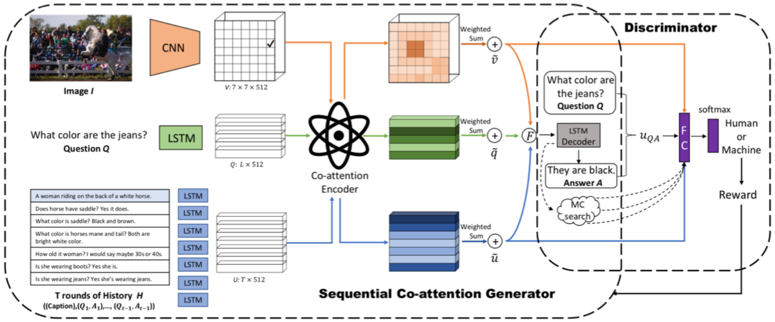
Related Work



▲ Late Fusion [1] (LF)



▲ Memory Network [1] (MN)

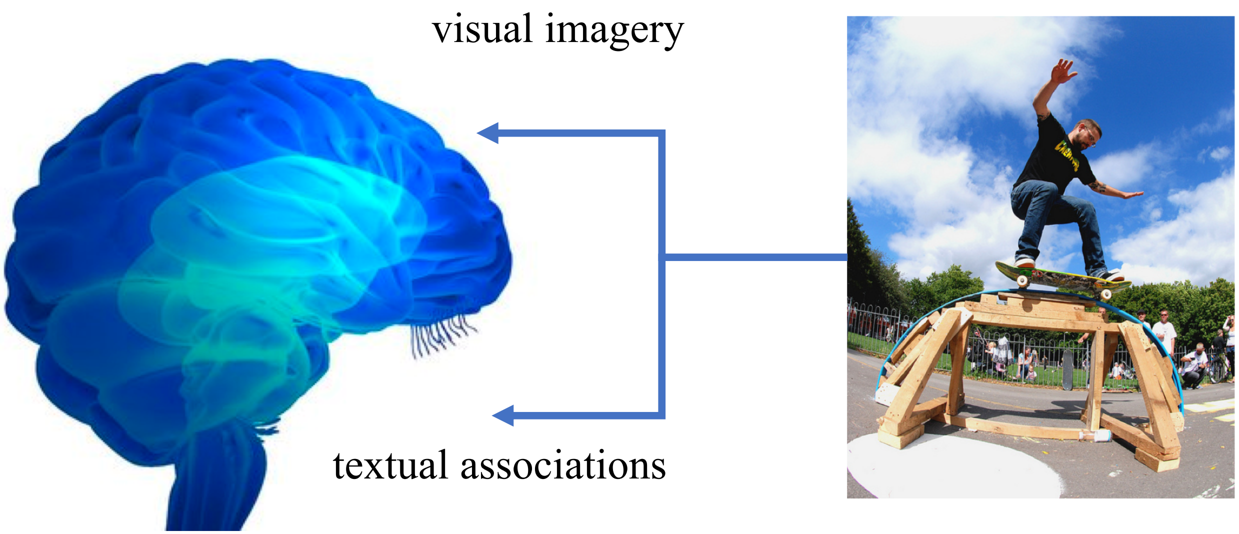


▲ Co-Attention[2] (CoAtt)

The role of visual information has been less studied！

[1] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *CVPR*, pages 1080–1089, 2017.
[2] Qi Wu, Peng Wang, Chunhua Shen, Ian Reid, and Anton van den Hengel. Are you talking tome? reasoned visual dialog generation through adversarial learning. In CVPR, pages 6106–6115, 2018.

Think from cognition view



Dual-coding theory [1] :

Our brain encodes information in two ways: visual imagery and textual associations.

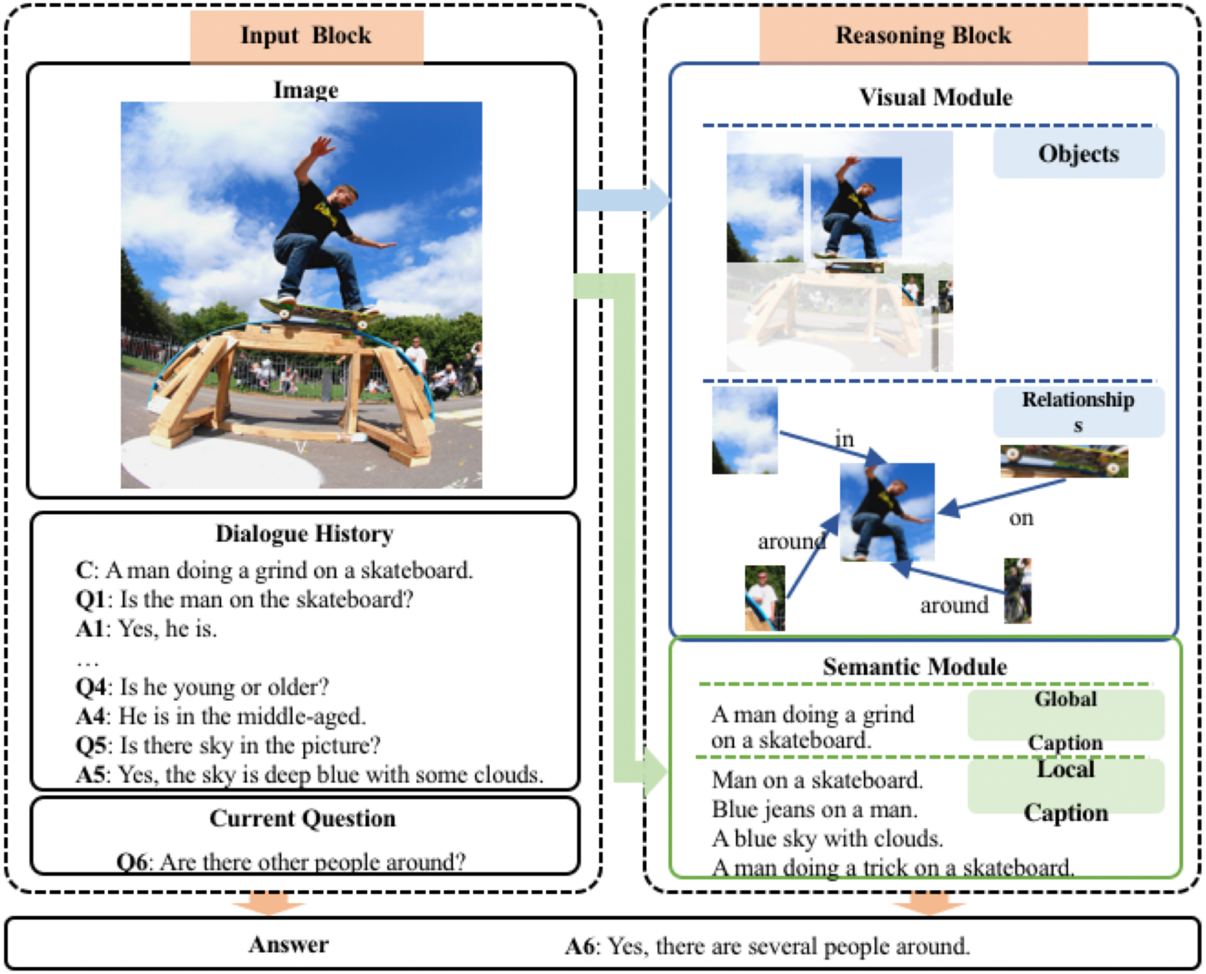
When asked to act upon a concept, our brain retrieves either images or words, or both simultaneously.

Encoding concept by two different ways strengthens the capacity of memory and understanding.

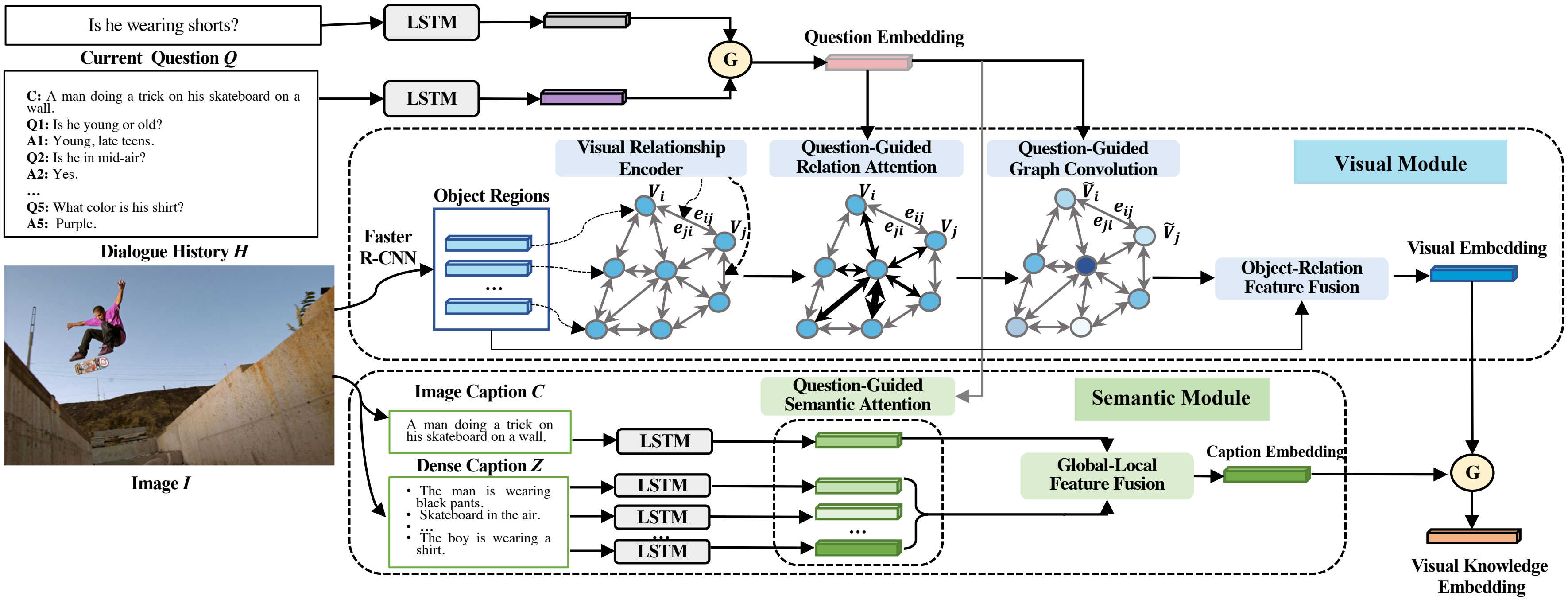
[1] A. Paivio, "*Imagery and Verbal Process*." New York: Holt, Rinehart and Winston., 1971.

DualVD

- Inspired by the cognitive process, we first propose a novel framework to comprehensively depict an image from both visual and semantic perspectives.



DualVD



- The core structure of the model is divided into two parts:
  Visual-Semantic Dual Encoding
  Adaptive Visual-Semantic Knowledge Selection

Experiments—SOTA Comparison

✓ Our model consistently outperforms all the approaches on most metrics and slightly underperforms the model using multi-step reasoning and complex attention mechanism.

TABLE I
RESULT COMPARISON ON VALIDATION SET OF VISDIAL V0.9.

Model	MRR	R@1	R@5	R@10	Mean
LF[7]	58.07	43.82	74.68	84.07	5.78
HRE[7]	58.46	44.67	74.50	84.22	5.72
HREA[7]	58.68	44.82	74.81	84.36	5.66
MN[7]	59.65	45.55	76.22	85.37	5.46
SAN-QI[17]	57.64	43.44	74.26	83.72	5.88
HieCoAtt-QI[42]	57.88	43.51	74.49	83.96	5.84
AMEM[43]	61.60	47.74	78.04	86.84	4.99
HCIAE[44]	62.22	48.48	78.75	87.59	4.81
SF[45]	62.42	48.55	78.96	87.75	4.70
CoAtt[10]	63.98	50.29	80.71	88.81	4.47
CorefMN[46]	**64.10**	**50.92**	80.18	88.81	4.45
VGNN[8]	62.85	48.95	79.65	88.36	4.57
DualVD-LF	62.94	48.64	80.89	89.94	4.17
DualVD-MN	63.12	48.89	**81.11**	**90.33**	**4.12**

TABLE II
RESULT COMPARISON ON TEST-STANDARD SET OF VISDIAL V1.0.

Model	MRR	R@1	R@5	R@10	Mean	NDCG
LF[7]	55.42	40.95	72.45	82.83	5.95	45.31
HRE[7]	54.16	39.93	70.47	81.50	6.41	45.46
MN[7]	55.49	40.98	72.30	83.30	5.92	47.50
LF-Att[7]	57.07	42.08	74.82	85.05	5.41	40.76
MN-Att[7]	56.90	42.43	74.00	84.35	5.59	49.58
CorefMN[46]	61.50	47.55	78.10	88.80	4.40	54.70
VGNN[8]	61.37	47.33	77.98	87.83	4.57	52.82
RvA[47]	63.03	49.03	80.40	89.83	4.18	55.59
DL-61[21]	62.20	47.90	80.43	89.95	4.17	**57.32**
DualVD-LF	63.23	49.25	80.23	89.70	4.11	56.32
DualVD-MN	**63.38**	**49.35**	**81.05**	**90.38**	**4.07**	57.09

TABLE III
RESULT COMPARISON ON VALIDATION SET OF VISDIAL-Q.

Model	MRR	R@1	R@5	R@10	Mean
LF[7]	18.45	7.80	26.12	40.78	20.42
MN[7]	39.83	25.80	54.76	69.80	9.68
SF-QI[39]	30.21	17.38	42.32	57.16	14.03
SF-QIH[39]	40.60	26.76	55.17	70.39	9.32
VGNN[8]	41.26	27.15	56.47	**71.97**	**8.86**
DualVD-LF	41.31	27.24	56.50	71.51	9.09
DualVD-MN	**41.34**	**27.27**	**56.60**	71.45	9.15

Experiments—Ablation Study

✓ Each component is effective.

Table 3: Ablation study of DualVD on VisDial v1.0.

Model	MRR	R@1	R@5	R@10	Mean	NDCG
ObjRep	63.84	49.83	81.27	90.29	4.07	55.48
RelRep	63.63	49.25	81.01	90.34	4.07	55.12
VisNoRel	63.97	49.87	81.74	90.60	4.00	56.73
VisMod	64.11	50.04	81.78	90.52	3.99	56.67
GlCap	60.02	45.34	77.66	87.27	4.78	50.04
LoCap	60.95	46.43	78.45	88.17	4.62	51.72
SemMod	61.07	46.69	78.56	88.09	4.59	51.10
DualVD	**64.64**	**50.74**	**82.10**	**91.00**	**3.91**	**57.30**

Experiments—Qualitative Analysis

- **Case**

Image | **Dialogue History**

C: 2 boys playing disc golf in a forest.

Question1	**Are the boys teenagers?**
Answer1	**They are young boys.**
Question2	**Do you see a lot of trees?**
Answer2	**Yes, a ton of trees.**
Question3	**Dose 1 of the boys holding the disc?**
Answer3	**They are both holding discs.**

Experiments—Qualitative Analysis

- The amount of information from each module highly depends on the complexity of the question and the relevance of the content.
- Simple questions about a single object depending more on the visual clues.

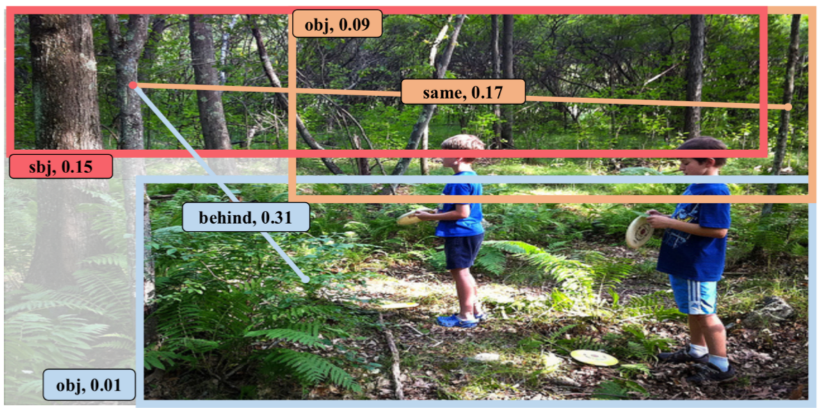
Question2	**Do you see a lot of trees?**
Answer2	**Yes, a ton of trees.**

Visual Module

Semantic Module

Ratio of total gate values: 60.20%

Ratio of total gate values: 39.80%



2 boys playing disc golf in a forest.

A man wearing blue shorts.

Boy holding blue frisbee.

Two people playing with a frisbee.

A blue shirt on a man.

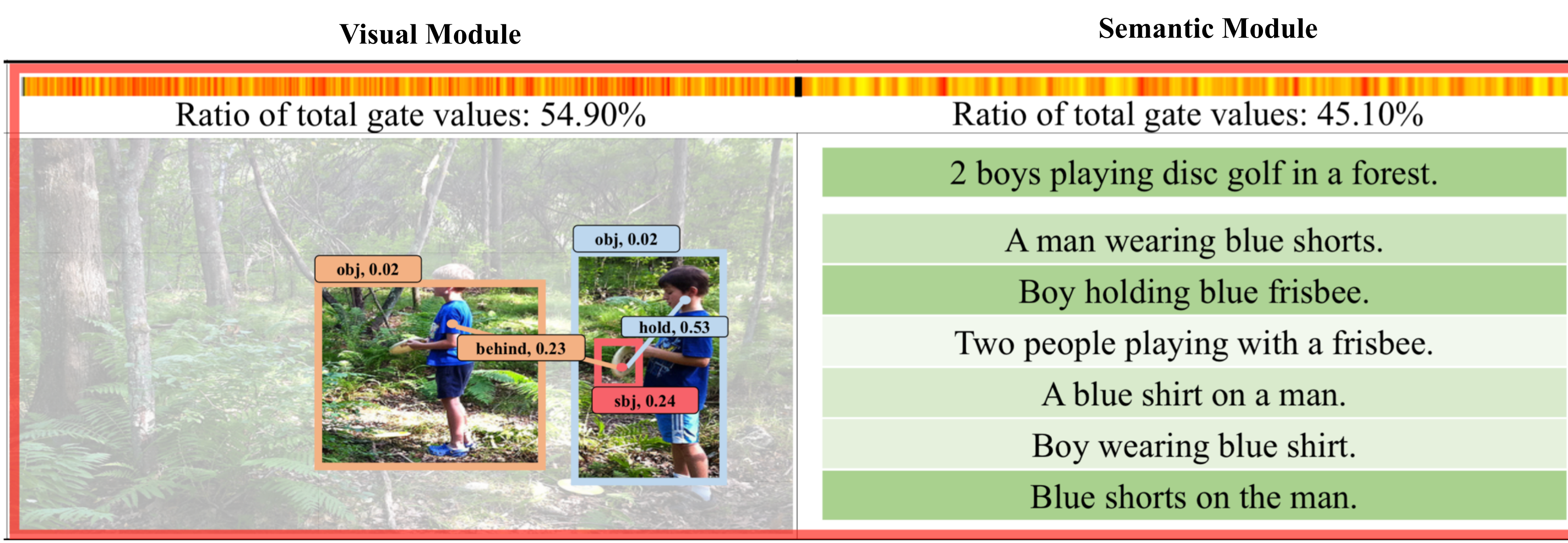
Boy wearing blue shirt.

Blue shorts on the man.

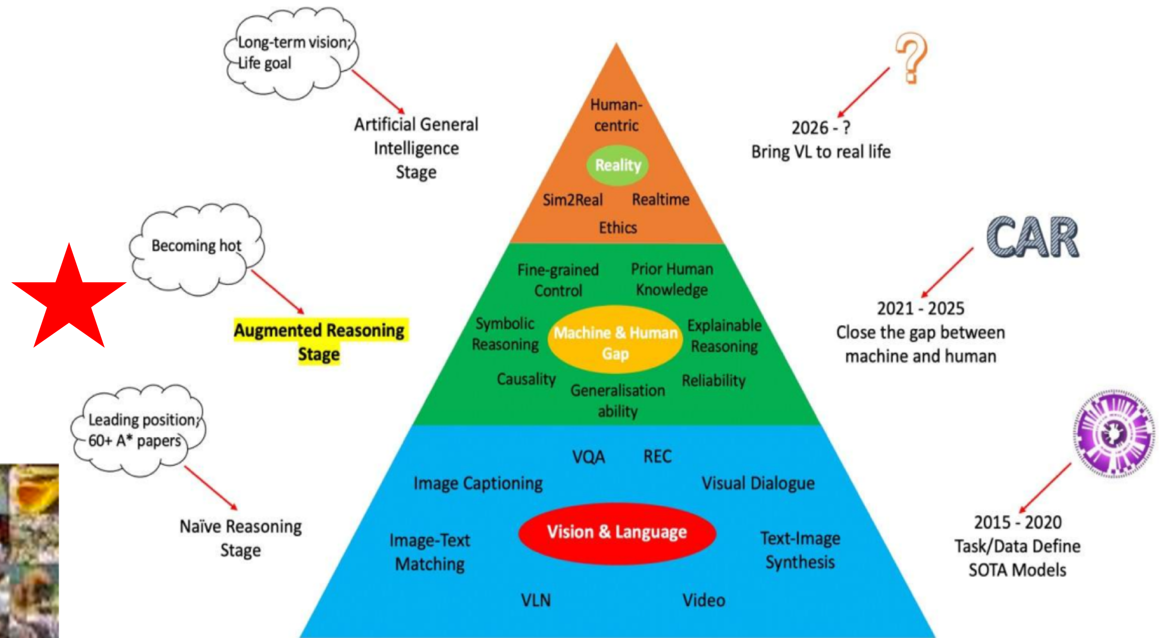
Experiments—Qualitative Analysis

- Complex questions about multiple objects and relationships require more semantic clues.
- Visual information is more important than semantic information to image understanding in visual dialogue.

Question3	**Dose 1 of the boys holding the disc?**
Answer3	**They are both holding discs.**



Take Home Message



Human Cognition
(Human Prior Knowledge)

Thanks! Q&A

Jing Yu

Email: yujing02@iie.ac.cn

Homepage: https://mmlab-iie.github.io/

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

中国科学院大学
University of Chinese Academy of Sciences