Microsoft

MuKEA: Multimodal Knowledge Extraction and Accumulation for Knowledge-based Visual Question Answering

CVPR 2022

Yang Ding , **Jing Yu*** , Bang Liu , Yue Hu, Mingxin Cui , Qi Wu

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

Université de Montréal

THE UNIVERSITY of ADELAIDE

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

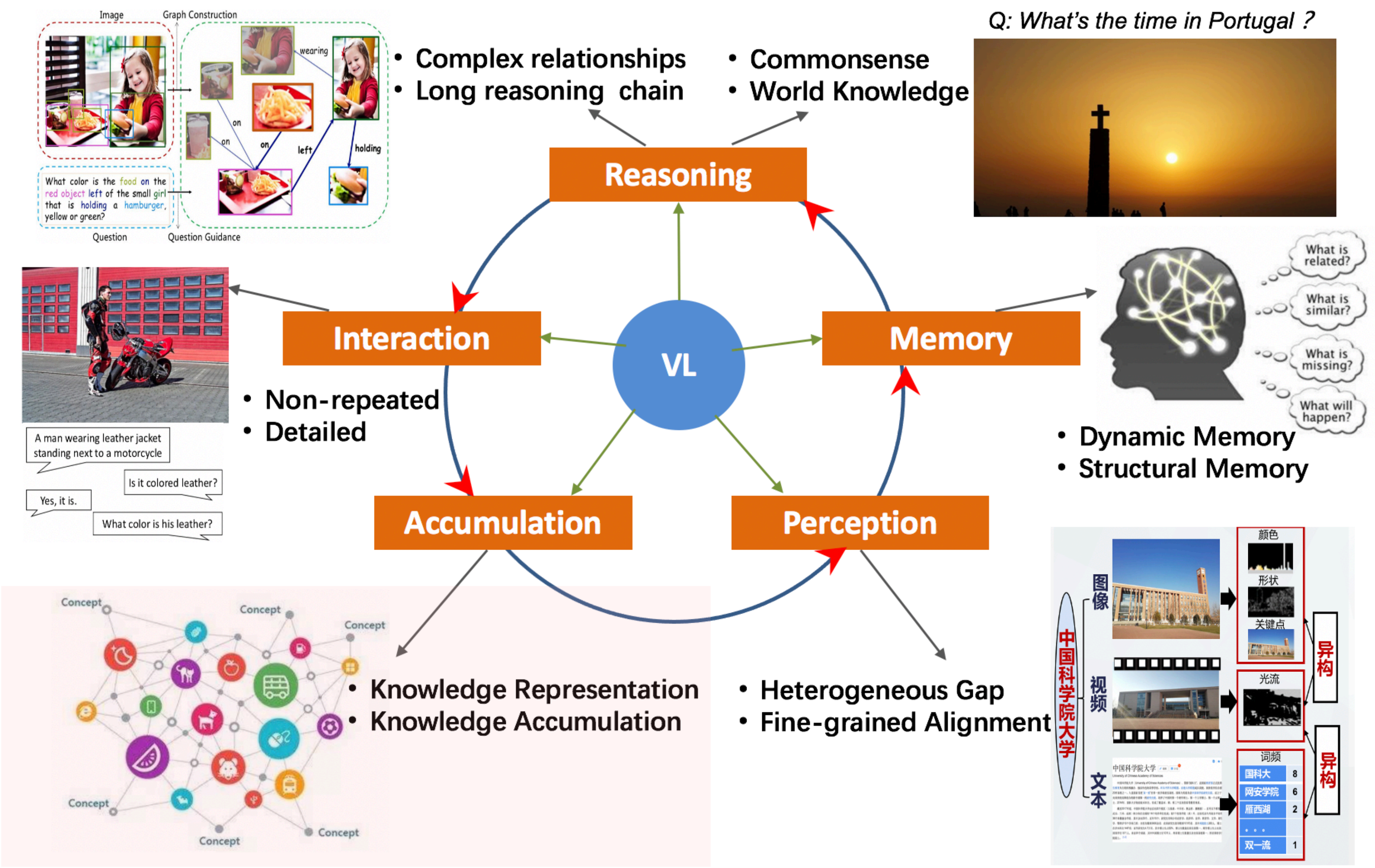
Mila

Microsoft

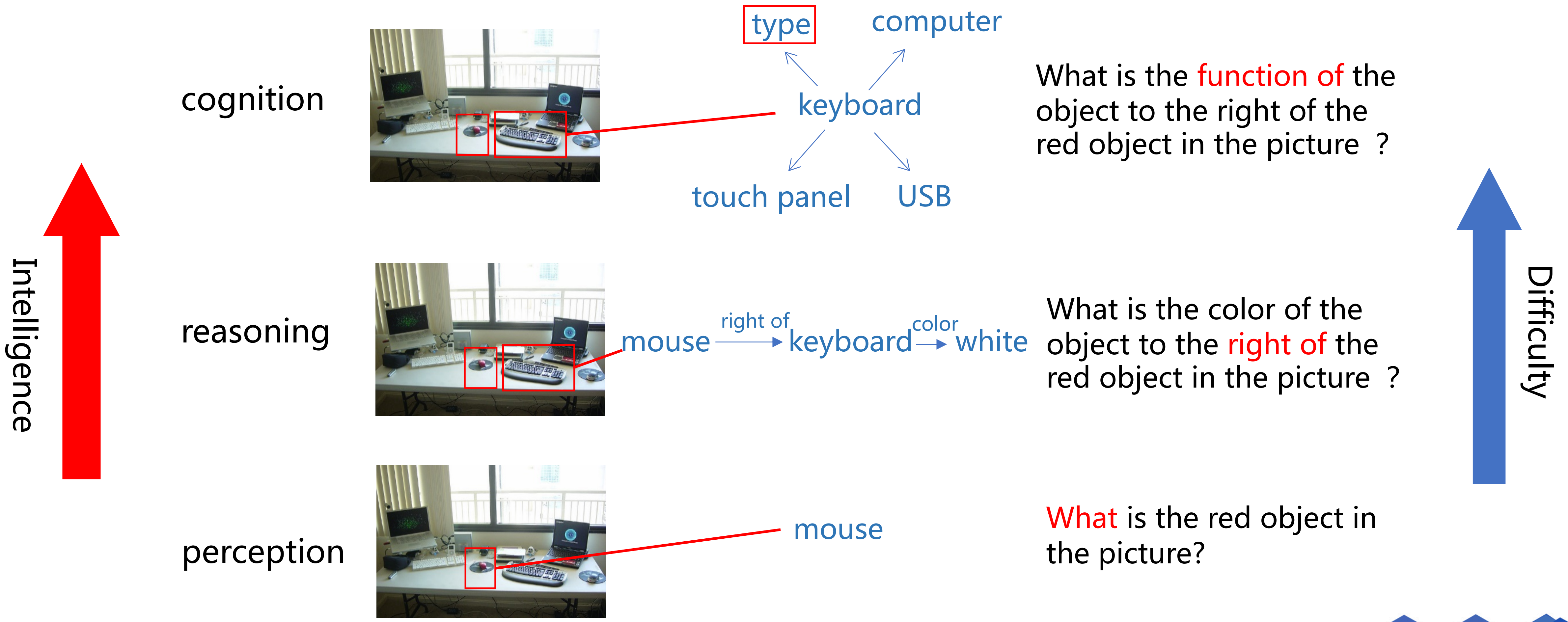
Content

- Motivation
- Model
- Experiments
- Summary

The roadmap of our CogModal group



- Visual question answering (VQA) evolves from perception to reasoning and then to cognition, requiring a gradually increase of intelligence.



Knowledge-based Visual Question Answering (KB-VQA) requires visual knowledge acquisition and reasoning.

Image

Question

What is the name of this type of motorcycle?

Knowledge Reasoning

Kawasaki

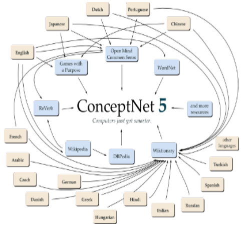
Answer

Unstructured Knowledge

WIKIPEDIA
The Free Encyclopedia

- OK-VQA [CVPR2020]
- Ask me anything [CVPR2016]
- Visual-Retriever-Reader [EMNLP2021]

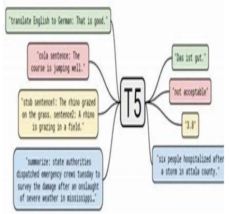
Structured Knowledge



DBpedia
Indonesian

- Conceptbert [EMNLP2020]
- Knowledge is power [SIGKDD2021]
- Mucko [IJCAI 2020]

Implicit Knowledge



- PICa [EMNLP2022]
- Frozen [NIPS2021]
- KAT [arXiv2022]

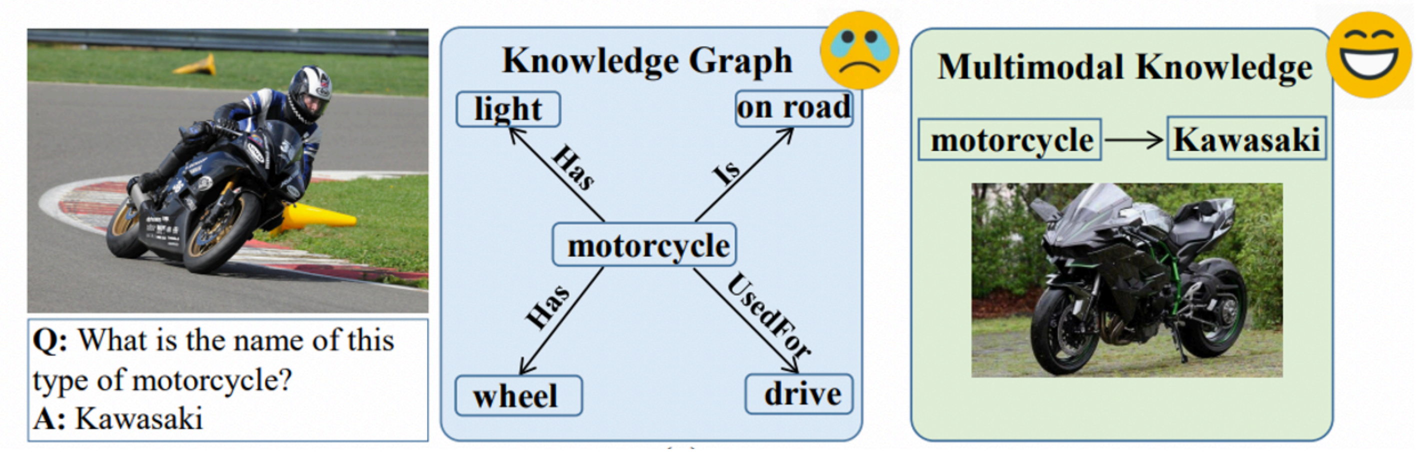
Multimodal Knowledge

motorcycle → **Kawasaki**

?

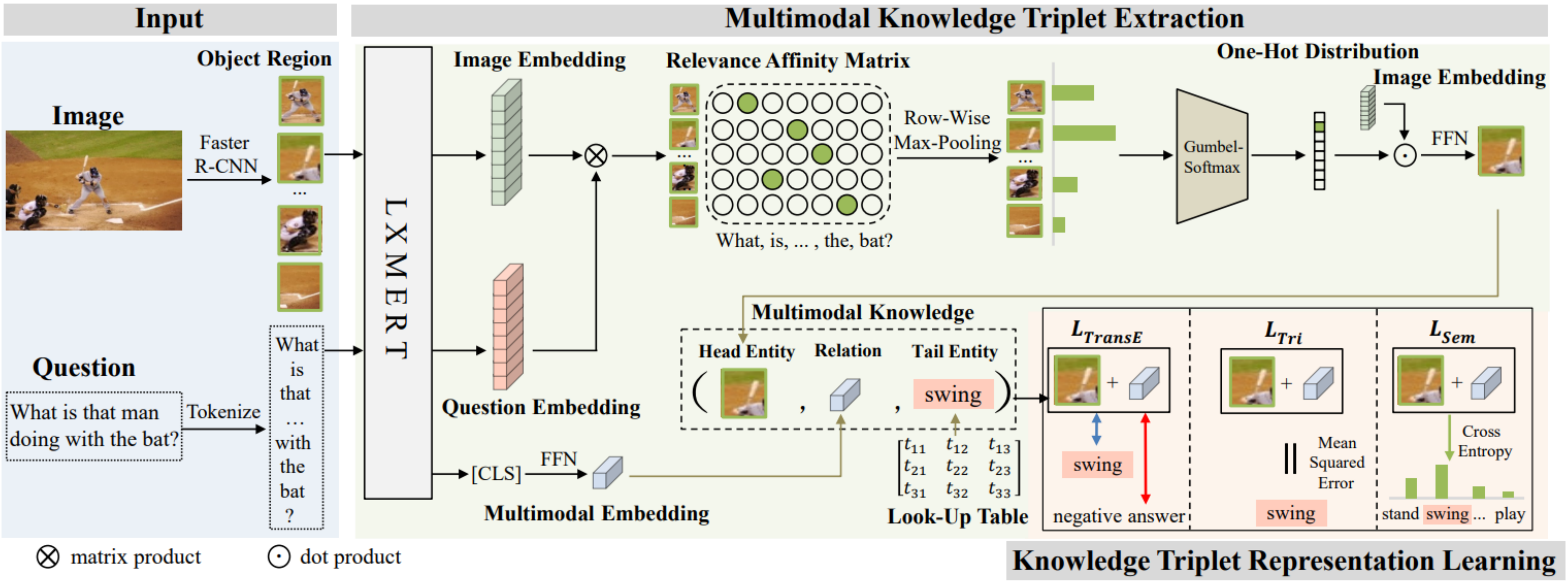
- KM⁴ [Information fusion2021]
- Gaia [ACL2020]
- MKGAT [CIKM2020]

Our Goal

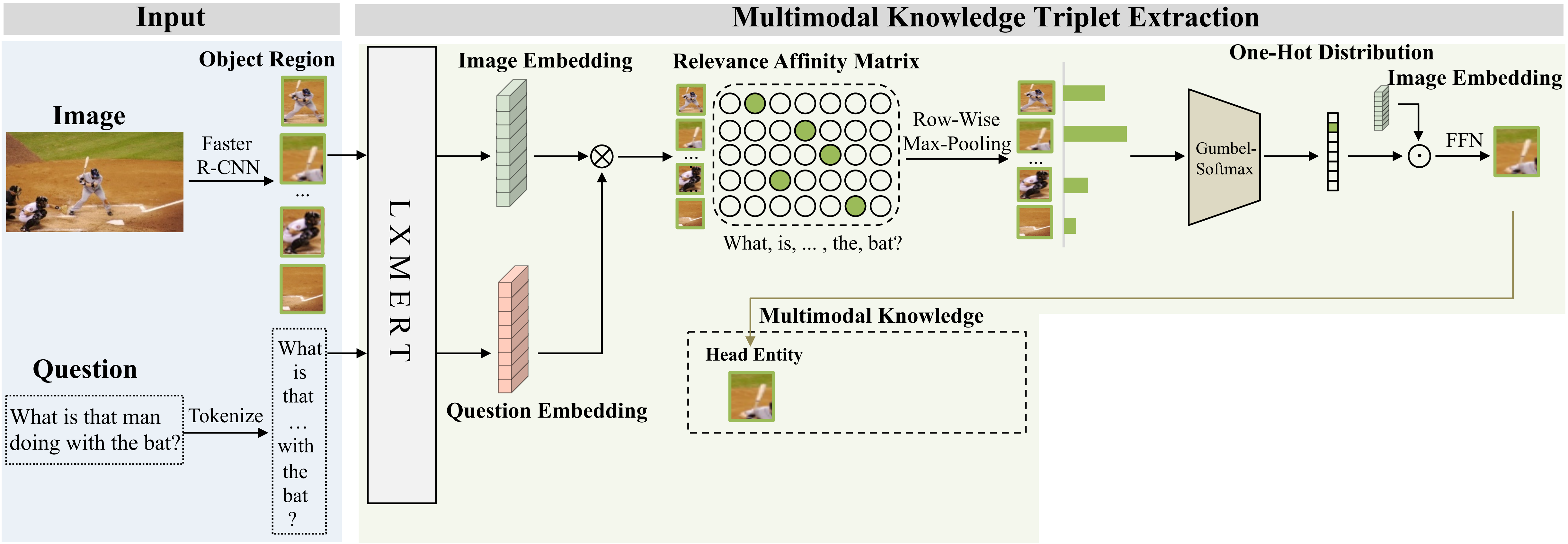


- How to represent the multimodal knowledge?
- How to accumulate the multimodal knowledge in the VQA scenarios?
- How to maintain the advantages of traditional knowledge graph in explainable reasoning?

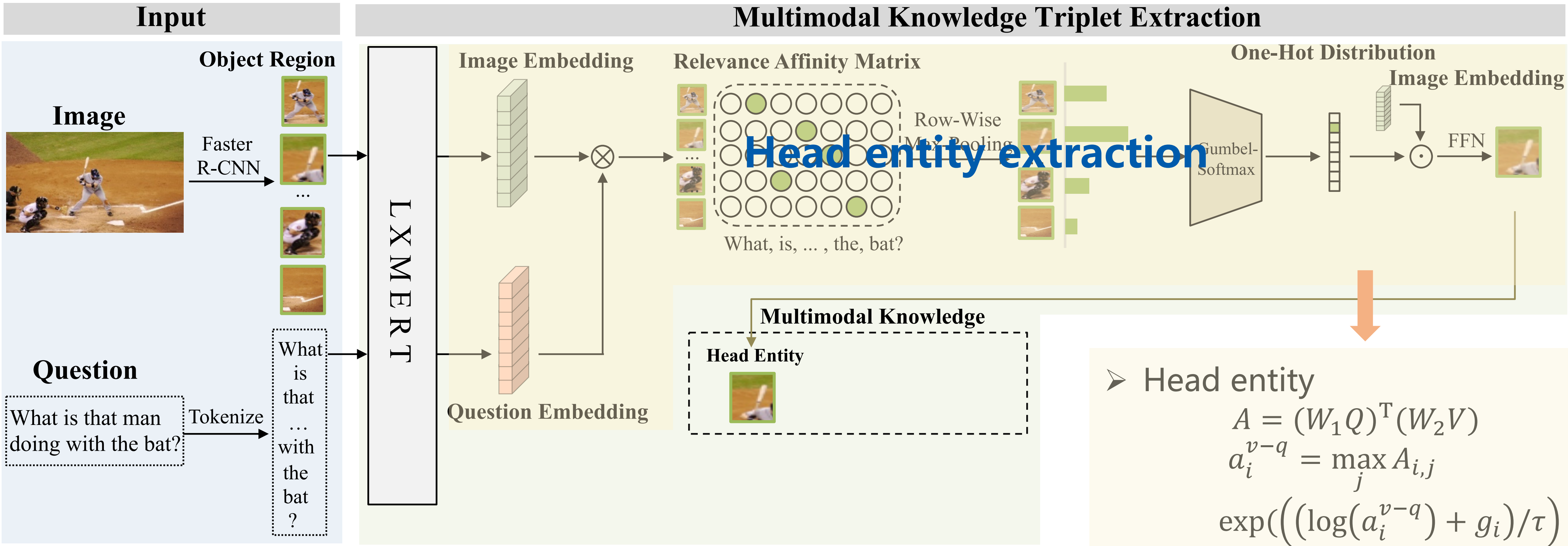
Model Framework



Multimodal Knowledge Triplet Extraction



Multimodal Knowledge Triplet Extraction



- Head entity

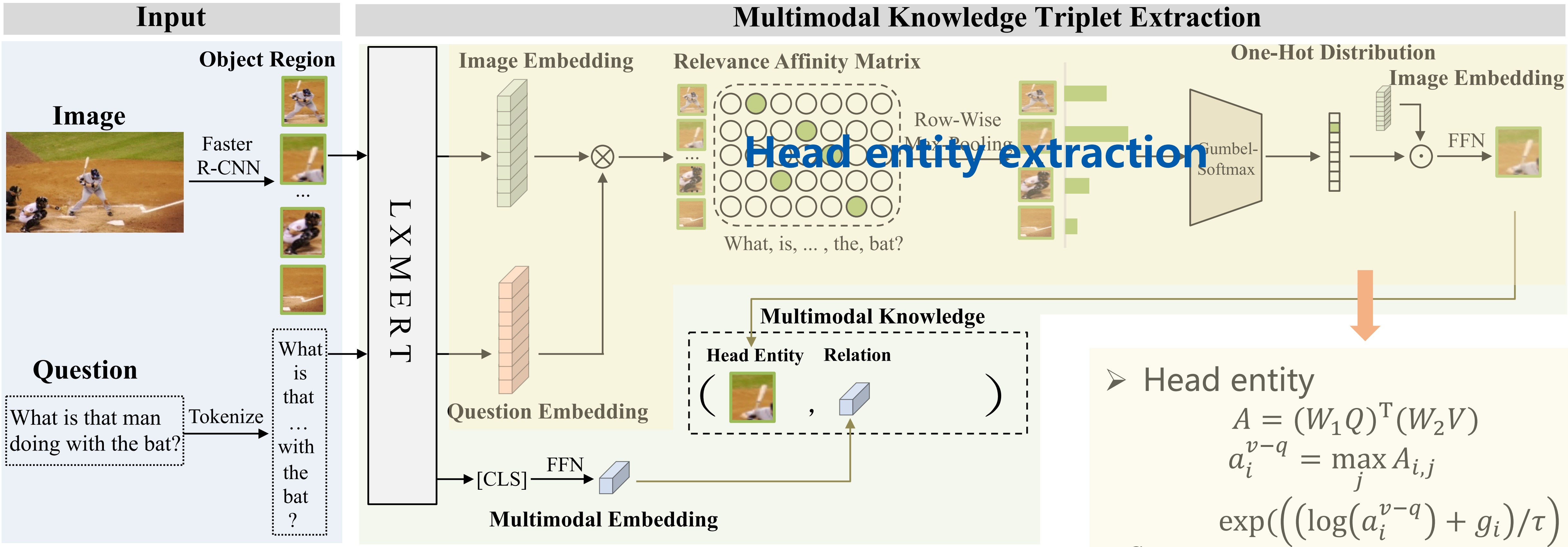
$$A = (W_1 Q)^{\mathrm{T}} (W_2 V)$$

$$a_i^{v-q} = \max_j A_{i,j}$$

$$a_i = \frac{\exp\left(\left(\log\left(a_i^{v-q}\right) + g_i\right)/\tau\right)}{\sum_{j=1}^{K} \exp\left(\left(\log\left(a_i^{v-q}\right) + g_i\right)/\tau\right)}$$

$$h = FFN\left(\sum_{i=1}^{K} a_i v_i\right)$$

Multimodal Knowledge Triplet Extraction



- Head entity

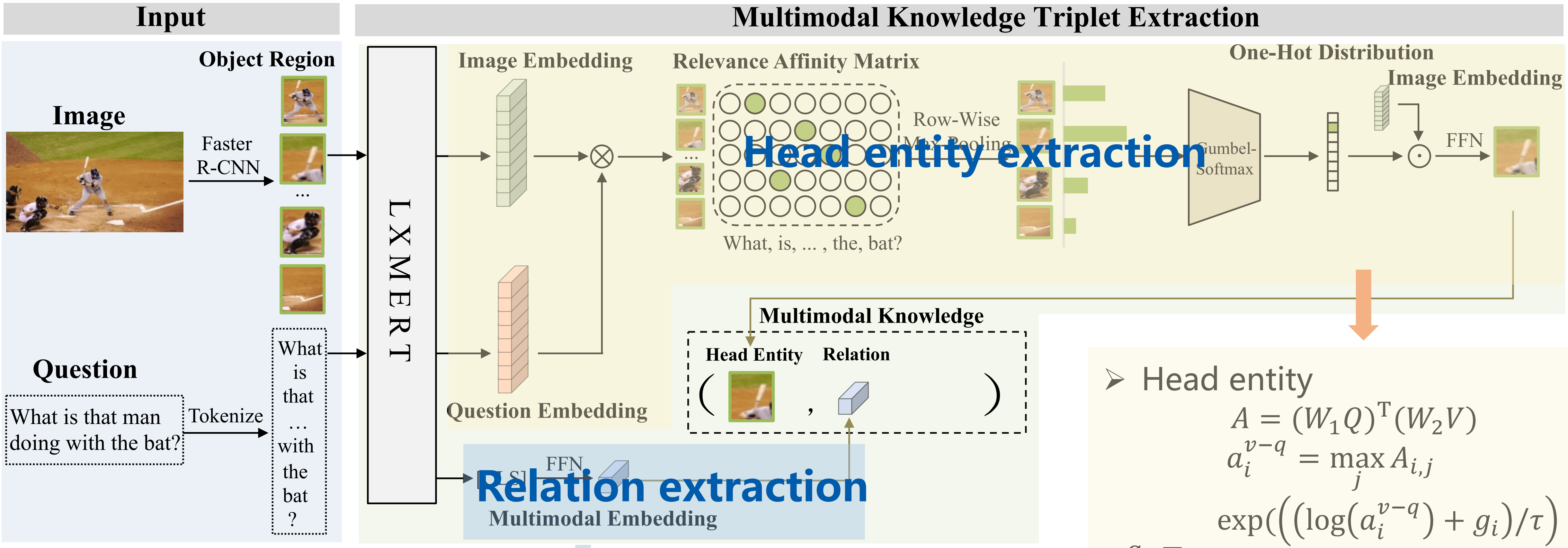
$$A = (W_1 Q)^{\mathrm{T}}(W_2 V)$$

$$a_i^{v-q} = \max_j A_{i,j}$$

$$a_i = \frac{\exp\left(\left(\log\left(a_i^{v-q}\right) + g_i\right)/\tau\right)}{\sum_{j=1}^{K} \exp\left(\left(\log\left(a_i^{v-q}\right) + g_i\right)/\tau\right)}$$

$$h = FFN\left(\sum_{i=1}^{K} a_i v_i\right)$$

Multimodal Knowledge Triplet Extraction



- Head entity

$$A = (W_1 Q)^{\mathrm{T}}(W_2 V)$$

$$a_i^{v-q} = \max_j A_{i,j}$$

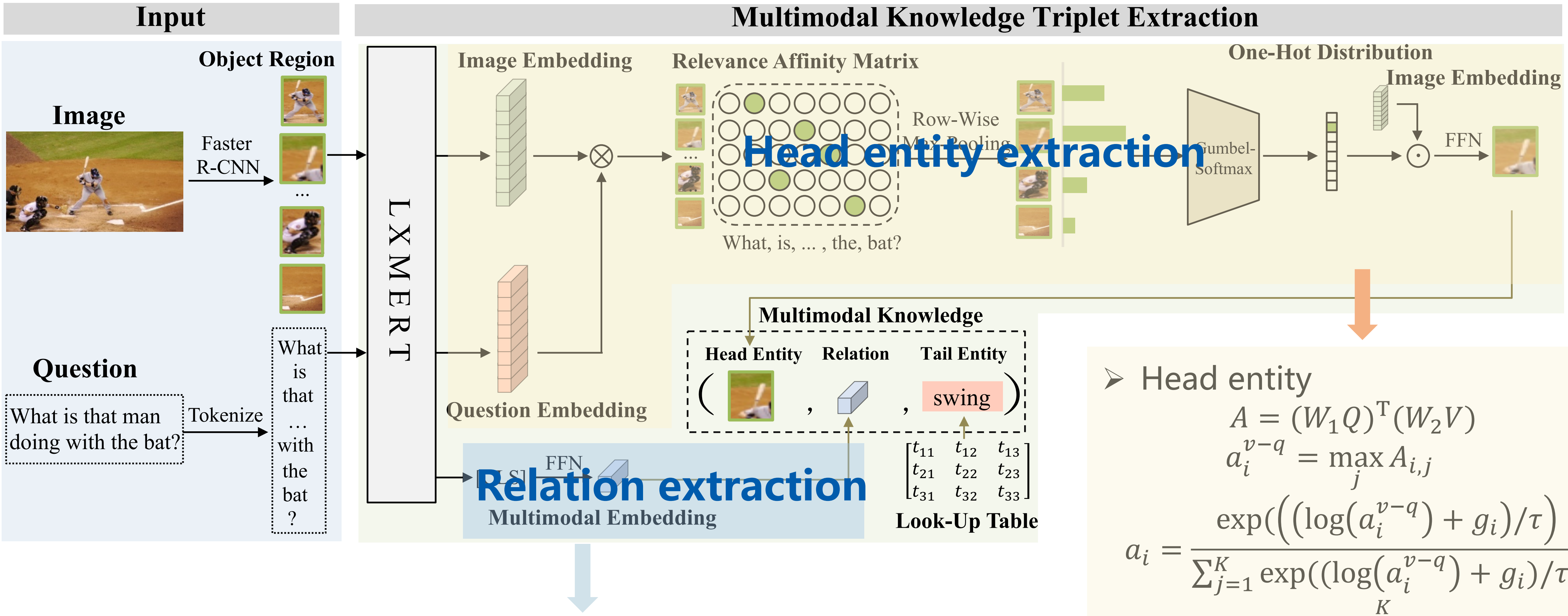
$$a_i = \frac{\exp\left(\left(\log\left(a_i^{v-q}\right) + g_i\right)/\tau\right)}{\sum_{j=1}^{K} \exp\left(\left(\log\left(a_i^{v-q}\right) + g_i\right)/\tau\right)}$$

$$h = FFN\left(\sum_{i=1}^{K} a_i v_i\right)$$

- Relation

$$r = FFN([CLS])$$

Multimodal Knowledge Triplet Extraction



- Relation

$$r = FFN([CLS])$$

- Head entity

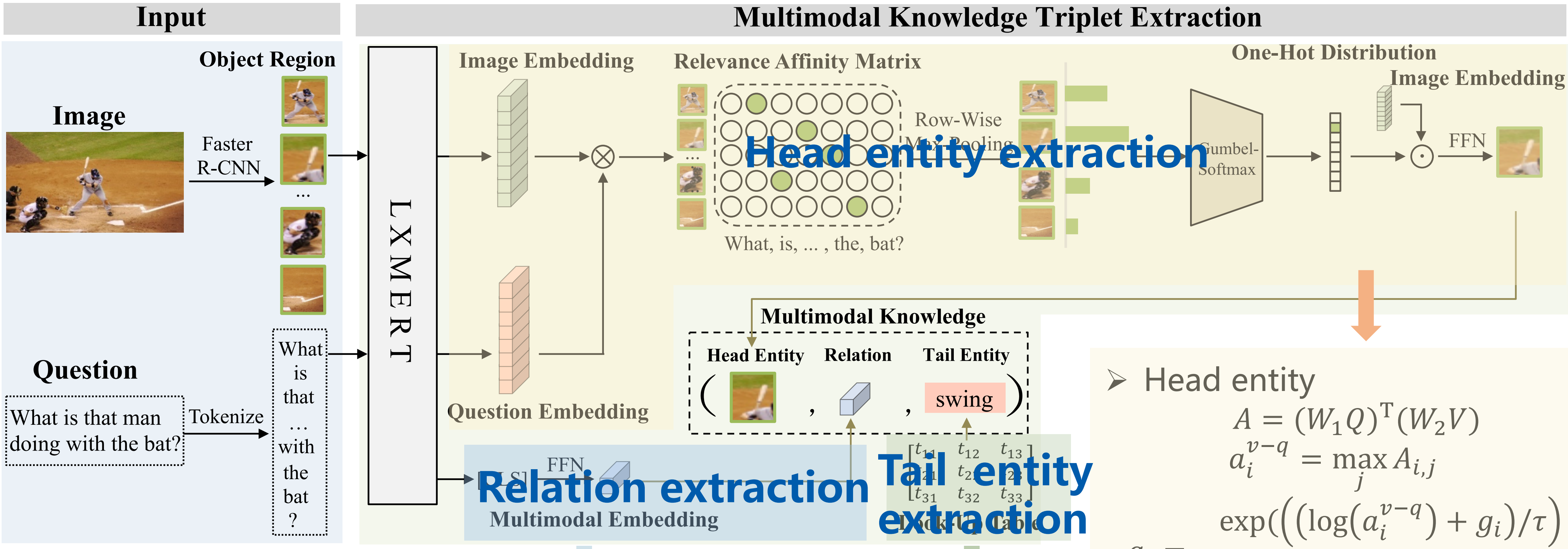
$$A = (W_1 Q)^{\mathrm{T}}(W_2 V)$$

$$a_i^{v-q} = \max_j A_{i,j}$$

$$a_i = \frac{\exp\left(\left(\log\left(a_i^{v-q}\right) + g_i\right)/\tau\right)}{\sum_{j=1}^{K} \exp\left(\left(\log\left(a_i^{v-q}\right) + g_i\right)/\tau\right)}$$

$$h = FFN\left(\sum_{i=1}^{K} a_i v_i\right)$$

Multimodal Knowledge Triplet Extraction



- Head entity

$$A = (W_1 Q)^{\mathrm{T}} (W_2 V)$$

$$a_i^{v-q} = \max_j A_{i,j}$$

$$a_i = \frac{\exp\left(\left(\log\left(a_i^{v-q}\right) + g_i\right)/\tau\right)}{\sum_{j=1}^{K} \exp\left(\left(\log\left(a_i^{v-q}\right) + g_i\right)/\tau\right)}$$

$$h = FFN\left(\sum_{i=1}^{K} a_i v_i\right)$$

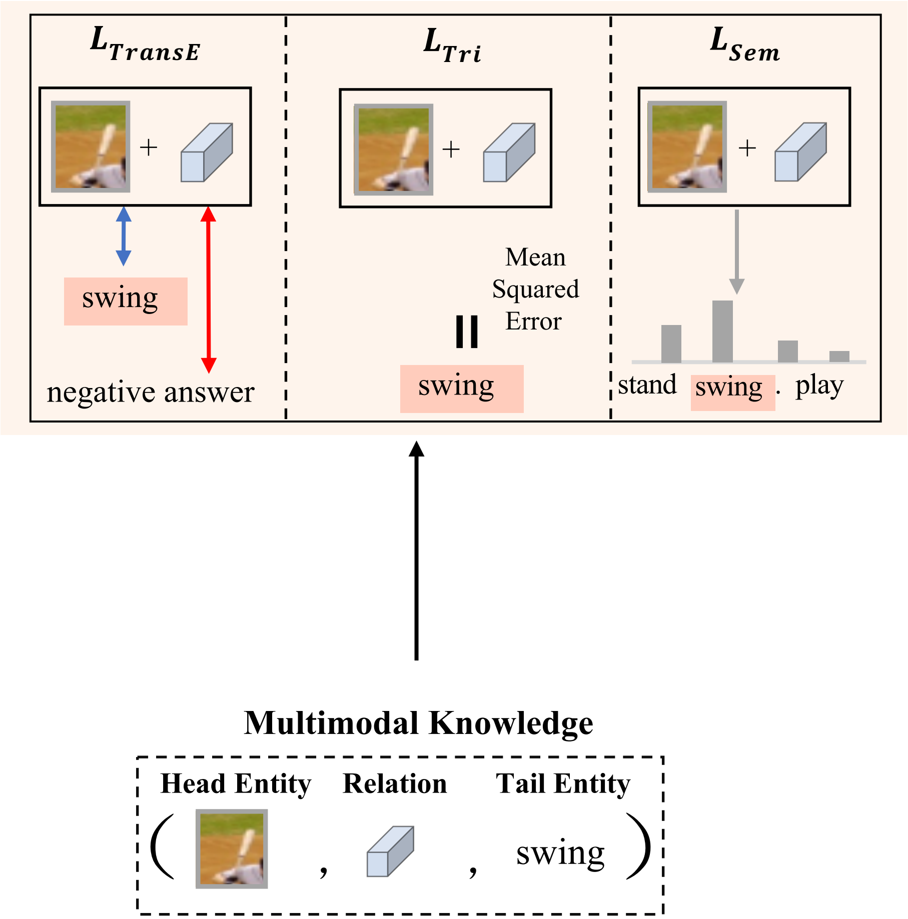
- Relation

$$r = FFN([CLS])$$

- Tail entity

Learning from scratch

Knowledge Triplet Representation Learning



- Preserve the embedding structure:

$$L_{TransE} = \sum_{t^+ \in A^+} \sum_{t^- \in A^-} [\gamma + d(h + r, t^+) - d(h + r, t^-)]_+$$

- Force the strict topological relation:

$$L_{Tri} = MSE(h + r, t^+)$$

- Learn a common semantic space:

$$P(t^+) = softmax\big((T)^T (h + r)\big)$$
$$L_{Sem} = -\log\big(P(t^+)\big)$$

- The final loss:

$$L = L_{TransE} + L_{Tri} + L_{Sem}$$

Knowledge Accumulation and Prediction

- **Pre-training**

VQA 2.0: basic visual dominant knowledge.

- **Fine-tuning**

OK-VQA/KR-VQA: more complex domain-specific multimodal knowledge.

- **Inference**

$$t_{in\ f} = arg \min_{t_i \in T} d(h_{in\ f} + r_{in\ f}\ , t_i)$$

Experiment Analysis

OK-VQA

Method	Knowledge Resources	Accuracy
ArticleNet (AN) [25]	Wikipedia	5.28
Q-only [25]	—	14.93
BAN [15]	—	25.17
+AN [25]	Wikipedia	25.61
+ KG-AUG [17]	Wikipedia + ConceptNet	26.71
MUTAN [5]	—	26.41
+ AN [25]	Wikipedia	27.84
Mucko [47]	ConceptNet	29.20
GRUC [42]	ConceptNet	29.87
KM^4 [45]	multimodal knowledge from OK-VQA	31.32
ViLBERT [21]	—	31.35
LXMERT [35]	—	32.04
KRISP(w/o mm pre.) [24]	DBpedia + ConceptNet + VisualGenome + haspartKB	32.31
KRISP(w/ mm pre.) [24]	DBpedia + ConceptNet + VisualGenome + haspartKB	38.90
ConceptBert [9]	ConceptNet	33.66
Knowledge is Power [46]	YAGO3	39.24
MuKEA	multimodal knowledge from VQA 2.0 and OK-VQA	**42.59**

- MuKEA achieves a remarkable boost of 3.35% on the overall metric over the best model
- End-to-end mode effectively avoids cascading error.
- MuKEA captures the question-centric and information-abstract multimodal knowledge

Experiment Analysis

KRVQA

Method	KB-not-related							KB-related					Overall
	one-step			two-step				one-step	two-step				
	0	1	2	3	4	5	6	2	3	4	5	6	
Q-type [7]	36.19	2.78	8.21	33.18	35.97	3.66	8.06	0.09	0.00	0.18	0.06	0.33	8.12
LSTM [7]	45.98	2.79	2.75	43.26	40.67	2.62	1.72	0.43	0.00	0.52	1.65	0.74	8.81
FiLM [30]	52.42	21.35	18.50	45.23	42.36	21.32	15.44	6.27	5.48	4.37	4.41	7.19	16.89
MFH [44]	43.74	28.28	27.49	38.71	36.48	20.77	21.01	12.97	5.10	6.05	5.02	14.38	19.55
UpDn [2]	56.42	29.89	28.63	49.69	43.87	24.71	21.28	11.07	8.16	7.09	5.37	13.97	21.85
MCAN [43]	49.60	27.67	25.76	39.69	37.92	21.22	18.63	12.28	9.35	9.22	5.23	13.34	20.52
+ knowledge retrieval [7]	51.32	27.14	25.69	41.23	38.86	23.25	21.15	13.59	**9.84**	9.24	5.51	13.89	21.30
MuKEA	**59.12**	**44.88**	**37.36**	**52.47**	**48.08**	**35.63**	**31.61**	**17.62**	6.14	**9.85**	**6.22**	**18.28**	**27.38**

- MuKEA consistently achieves a remarkable boost of 6.08% on the overall metric over the best model
- Even the vision-only questions require multimodal commonsense to bridge the low-level visual content and high-level semantics.

Experiment Analysis

Ablation Study

Method	Accuracy
1. MuKEA (full model)	**42.59**
Ablation of Loss Function	
2. w/o $\mathcal{L}_{\mathrm{Tri}}$	41.35
3. w/o $\mathcal{L}_{\mathrm{Sem}}$	42.06
4. w/o $\mathcal{L}_{\mathrm{Tri}}$ & $\mathcal{L}_{\mathrm{Sem}}$	40.84
5. w/o $\mathcal{L}_{\mathrm{TransE}}$	24.50
Ablation of Triplet Representation	
6. head entity w/ soft-attention	40.67
7. relation w/ self-attention	40.79
8. tail entity w/ GloVe	41.42
Ablation of Triplet Structure	
9. w/o h	39.83
10. w/o r	39.40
Ablation of Knowledge Source	
11. w/o VQA 2.0 knowledge	36.35
12. w/o OK-VQA knowledge	27.20
Ablation of Pre-training Knowledge	
13. w/o LXMERT pre-training	33.52

- Confirm the complementary of each loss function.
- Assess the influence of triplet extraction methods.
- Prove the importance of triplet structure.
- Both basic knowledge and domain-specific knowledge are important.
- Influence of prior knowledge accumulated in the pre-trained LXMERT

Experiment Analysis

Knowledge Complementary Analysis

Method	Failure subset		
	MUTAN + AN*	Mucko*	KRISP*
MuKEA	40.09	40.06	40.46

(a)

Method	Failure subset
	MuKEA
MUTAN + AN*	26.45
Mucko*	27.68
KRISP*	27.68

(b)

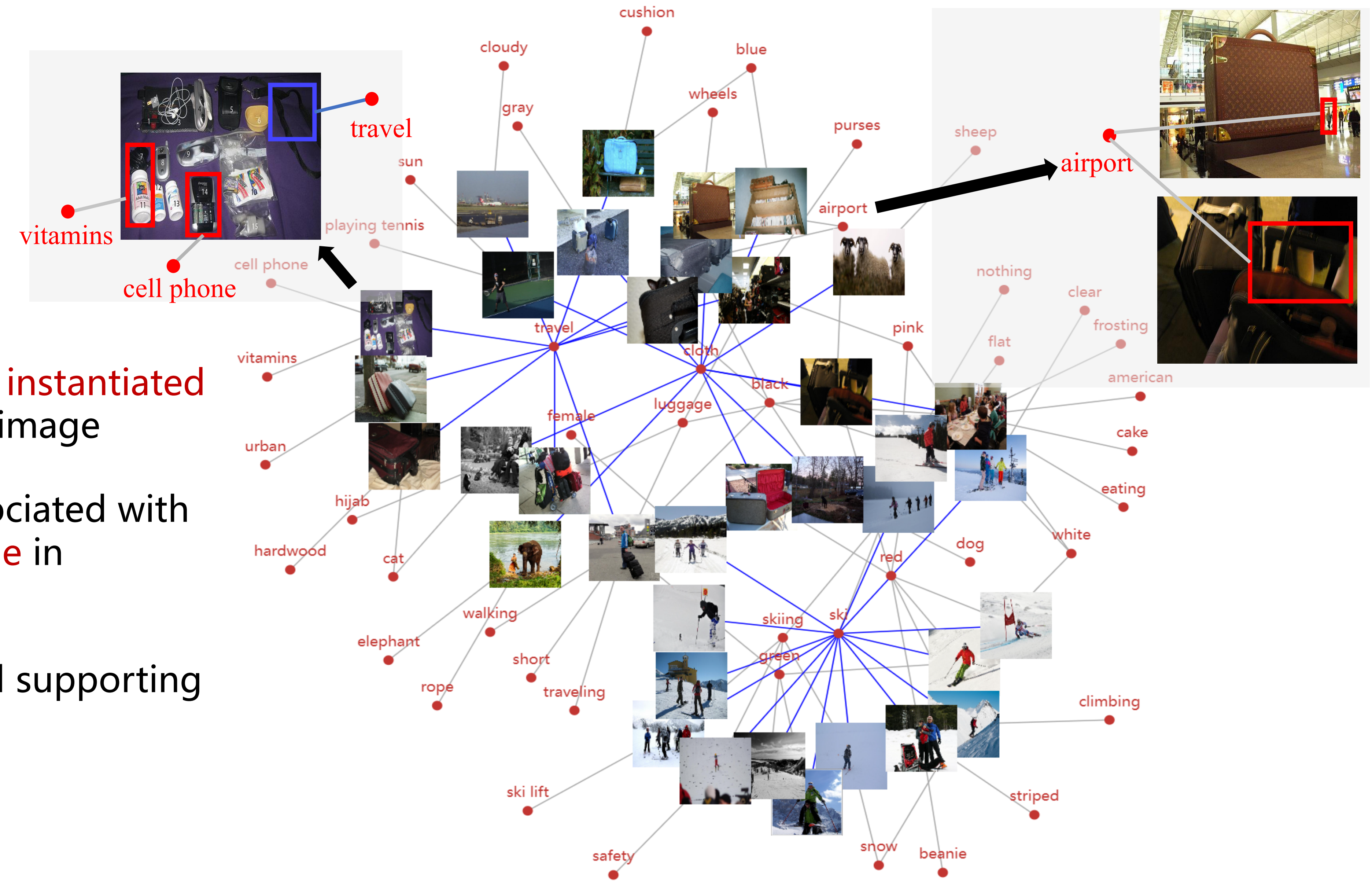
Method	Accuracy
MuKEA	42.59
MUTAN + AN*	25.43
MuKEA + (MUTAN + AN*)	35.39
MuKEA + (MUTAN + AN*) oracle	43.64
Mucko*	27.17
MuKEA + Mucko*	35.97
MuKEA + Mucko* oracle	44.84
KRISP*	32.02
MuKEA + KRISP*	37.75
MuKEA + KRISP* oracle	47.15

- Multimodal knowledge and existing KB knowledge respectively deals with different types of open-ended question

- Complementary benefits of multimodal knowledge and existing knowledge bases

Experiment Analysis

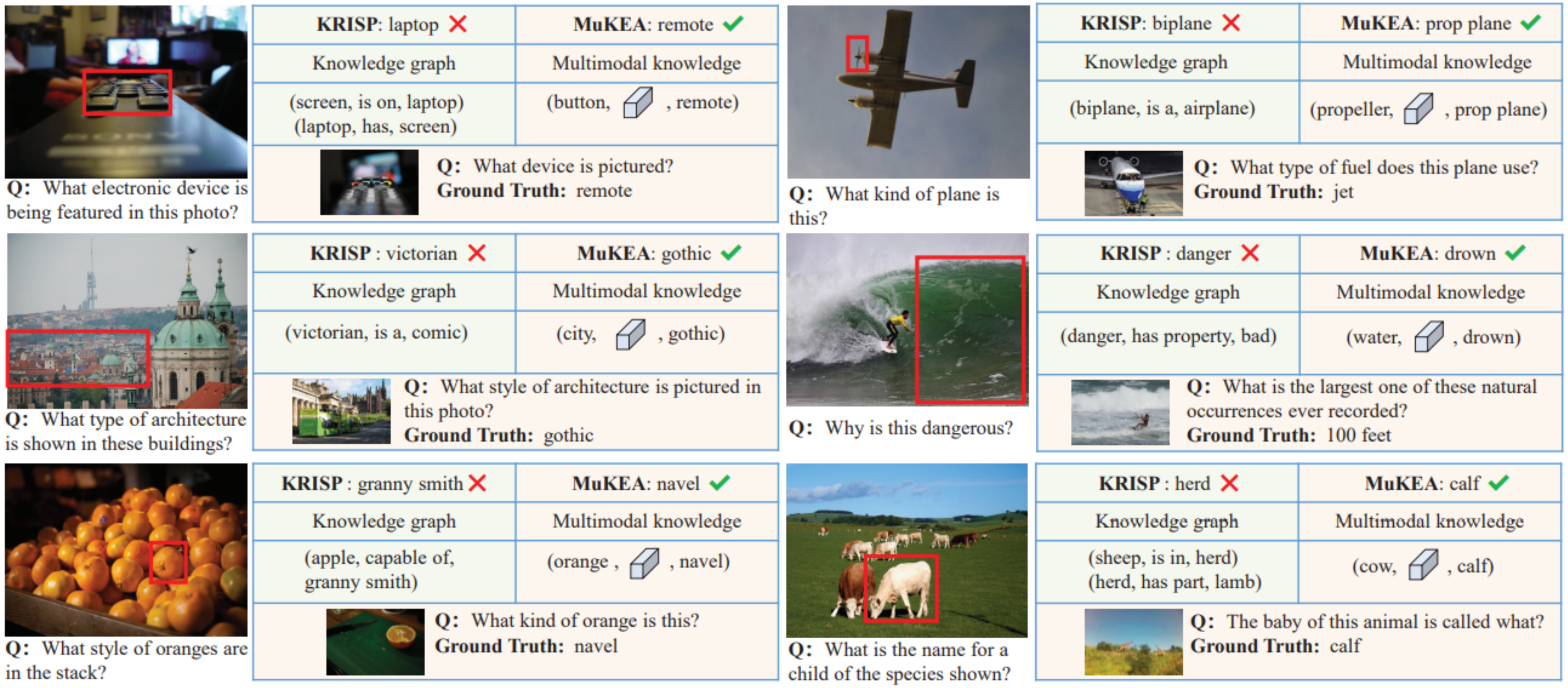
Accumulated Multimodal Knowledge



- MuKEA extracts different instantiated knowledge for the same image
- The same concept is associated with different visual knowledge in different scenes.
- Relation is extensible and supporting retrieval.

Experiment Analysis

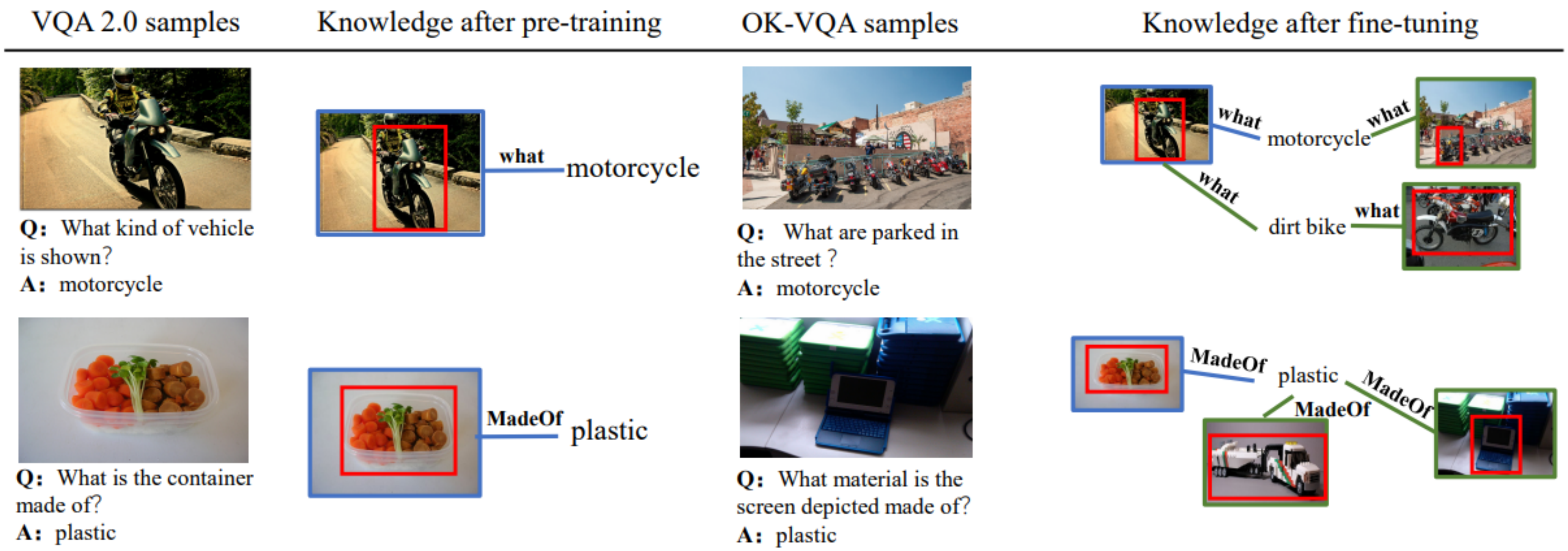
The Predicted Multimodal Knowledge Triplets



- MuKEA captures instantiated knowledge
- MuKEA contains multi-object involved complex knowledge
- MuKEA avoids the cascading error.

Experiment Analysis

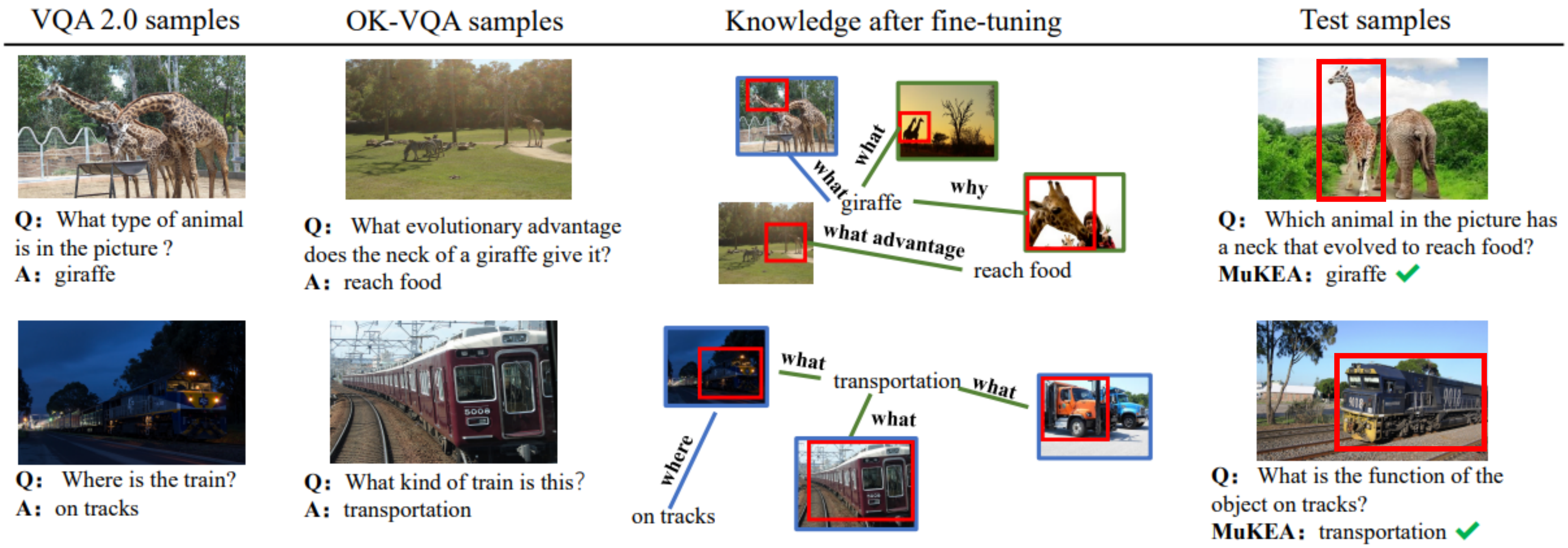
Progressive Knowledge Accumulation



- We illustrate how the basic visual knowledge in VQA 2.0 helps to learn more complex knowledge in OK-VQA.

Experiment Analysis

Zero-shot Analysis of Accumulated Multi-modal Knowledge



- MuKEA correlates ‘giraffe’ with ‘evolution’ through the manually constructed question.

Summary and Future Work

➢ Summary

- MuKEA focuses on multimodal knowledge instead of language knowledge for KB-VQA.
- Multimodal knowledge is represented by explicit triplets via three loss functions.
- A pre-training and fine-tuning strategy accumulates multimodal knowledge from basic to complex.

➢ Future Work

- How to effectively combine multimodal knowledge with existing knowledge bases?
- How to accumulate generic multimodal knowledge for vision-language tasks?

Thanks! Q&A

Jing Yu

Email: yujing02@iie.ac.cn

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

Homepage

Paper

Code

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