



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



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

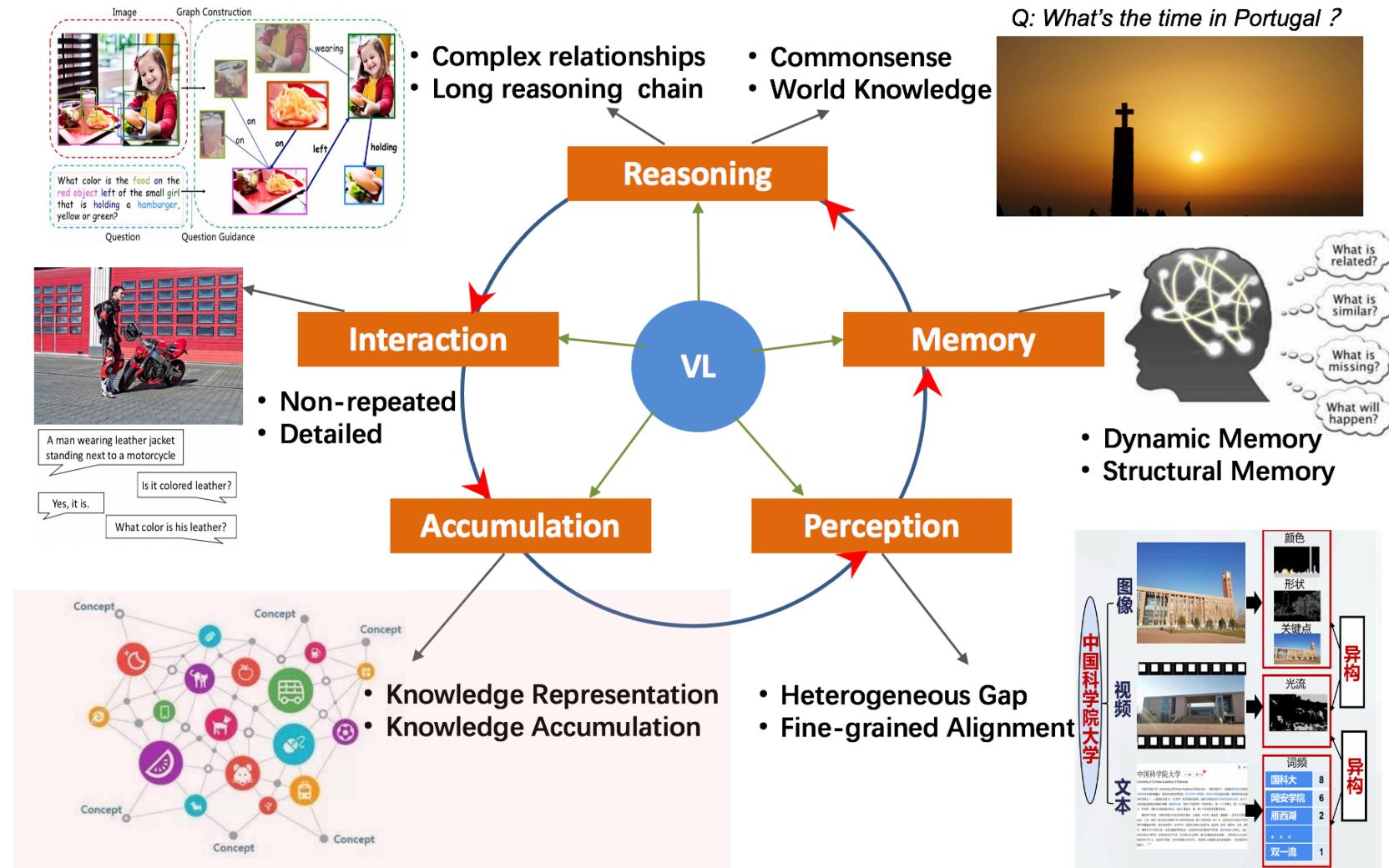


Content

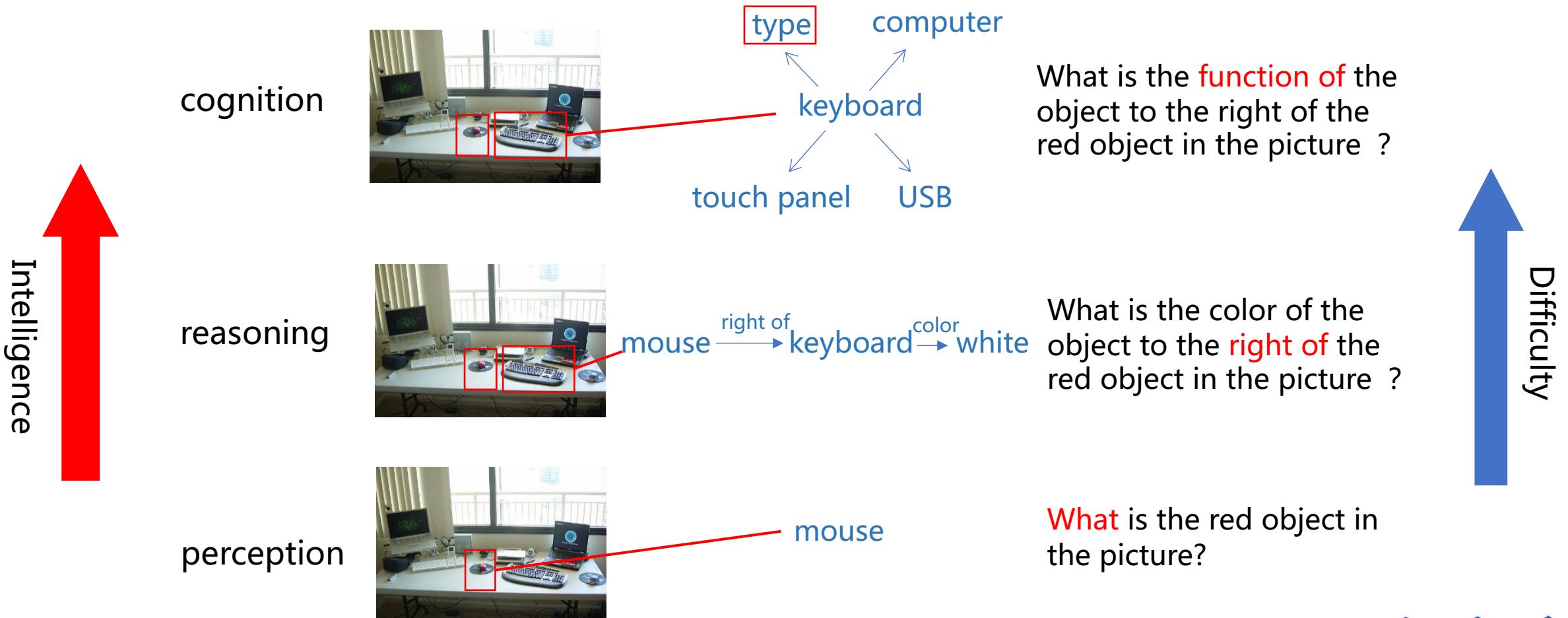
- Motivation
- Model
- Experiments
- Summary



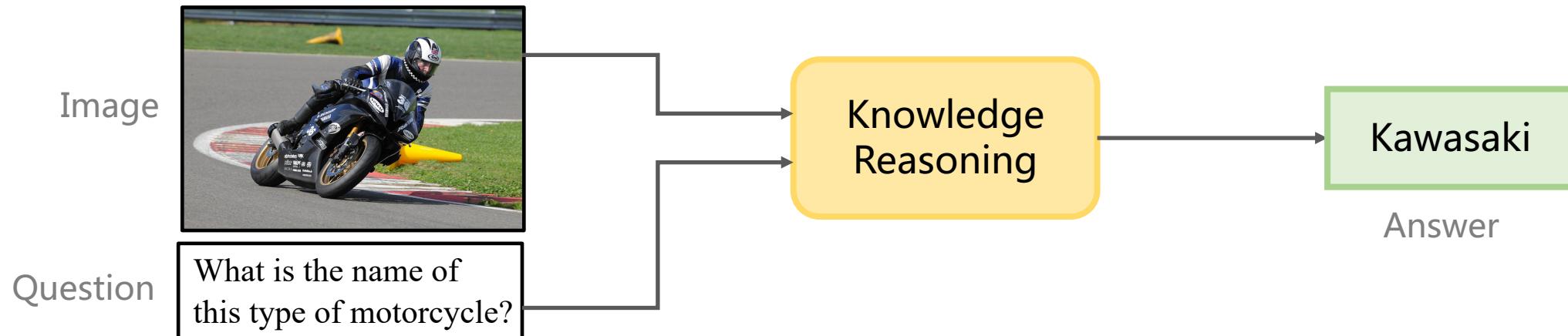
The roadmap of our CogModal group



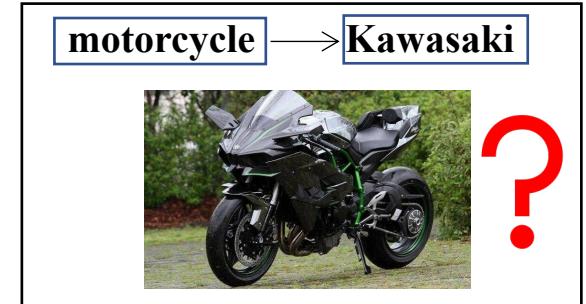
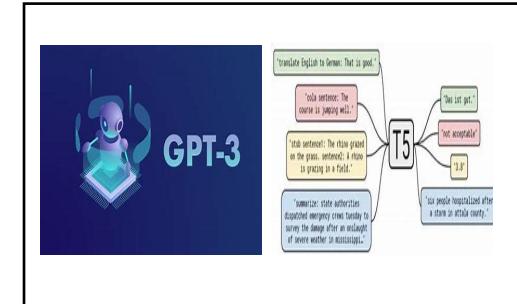
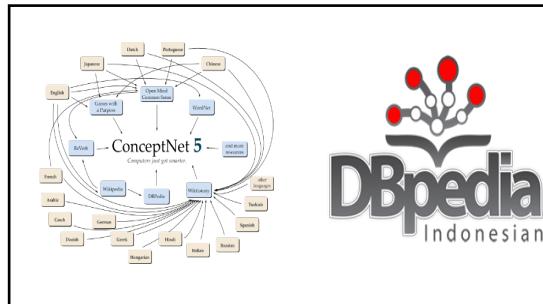
- ◆ Visual question answering (VQA) evolves from perception to reasoning and then to cognition, requiring a gradually increase of intelligence.



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



Unstructured Knowledge Structured Knowledge Implicit Knowledge Multimodal Knowledge



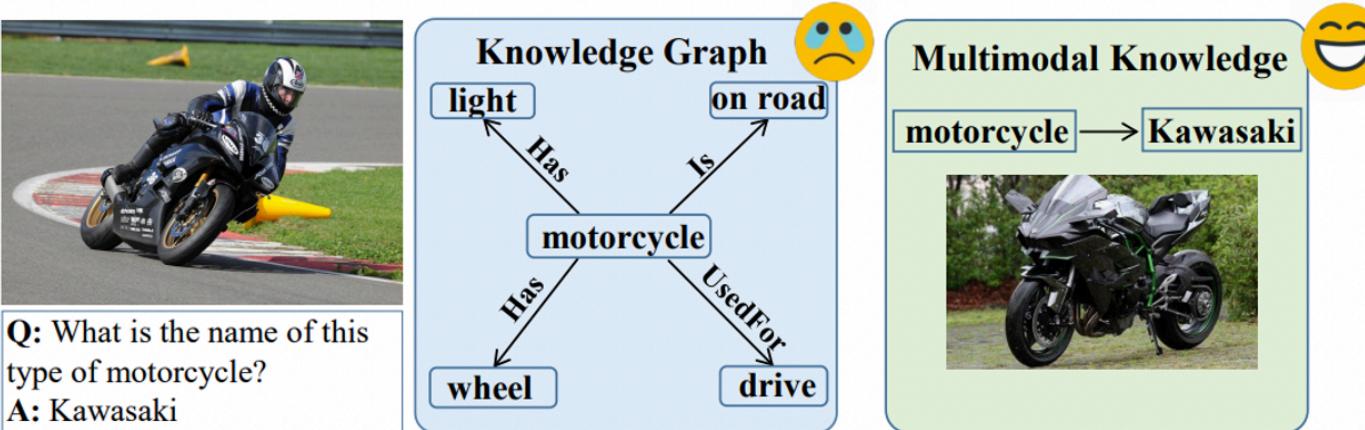
- OK-VQA [CVPR2020]
- Ask me anything [CVPR2016]
- Visual-Retriever-Reader [EMNLP2021]

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

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

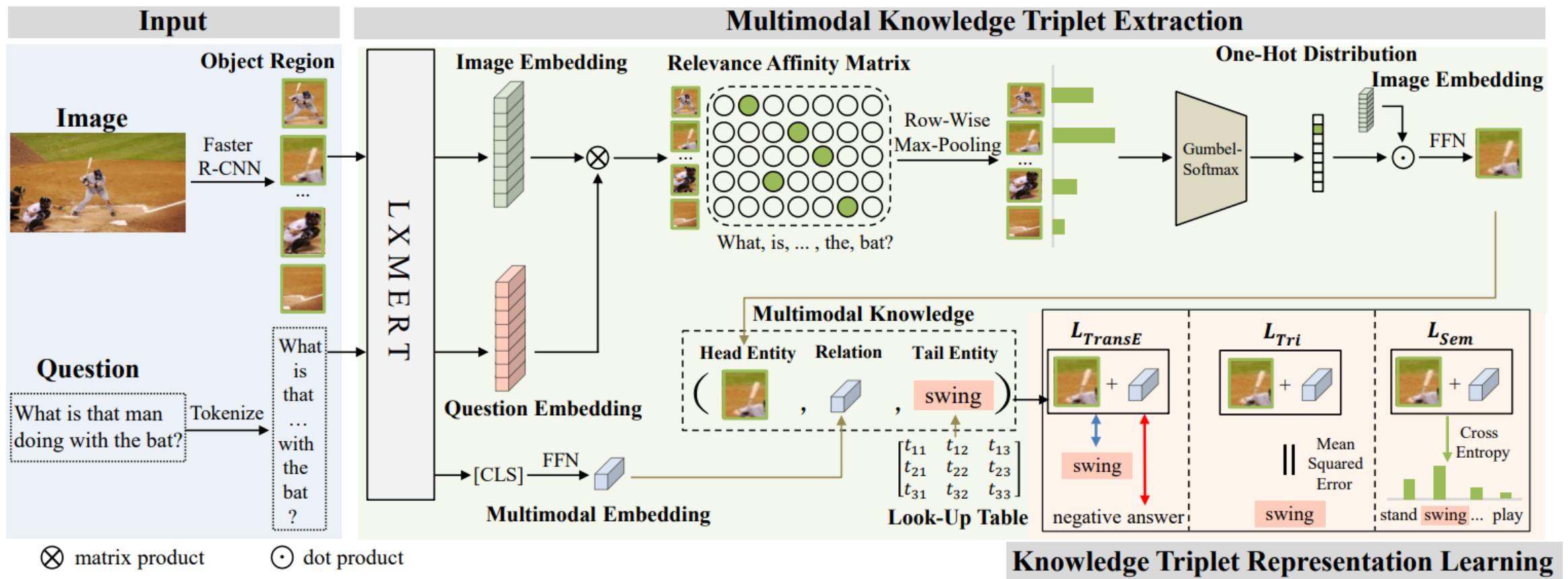
- KM⁴ [Information fusion2021]
- Gaia [ACL2020]
- MKGAT [CIKM2020]

Our Goal

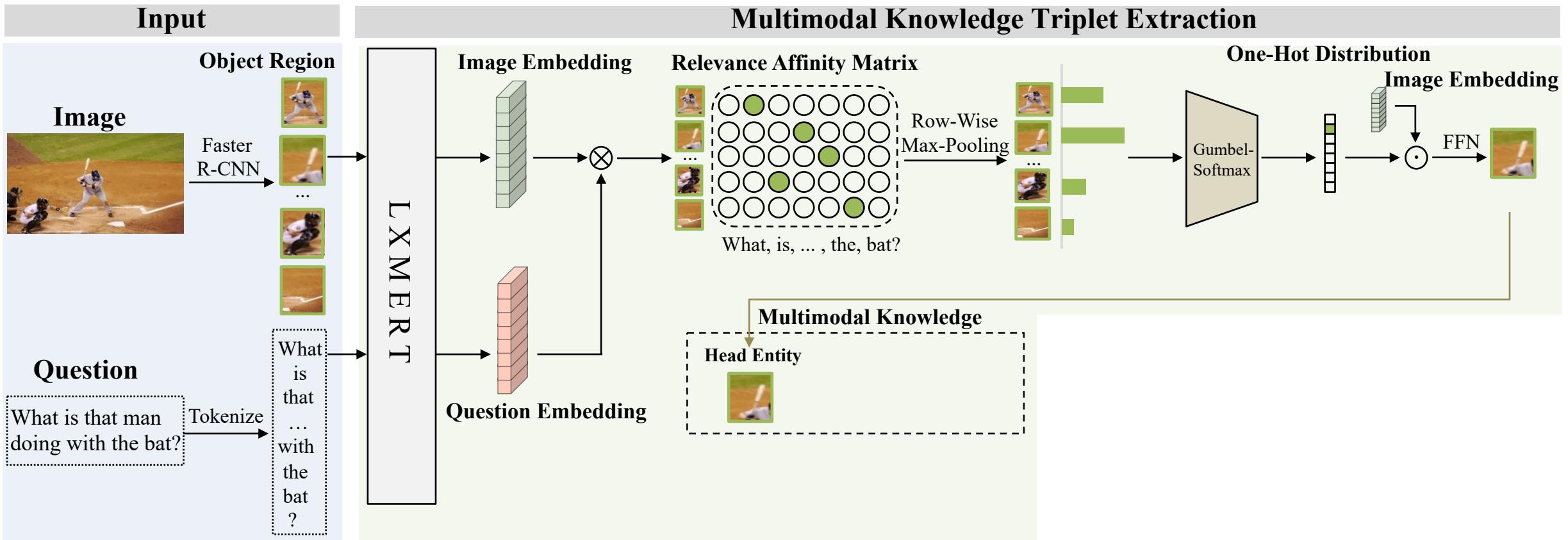


- How to **represent** the multimodal knowledge?
- How to **accumulate** the multimodal knowledge in the VQA scenarios?
- How to maintain the advantages of traditional knowledge graph in **explainable reasoning**?

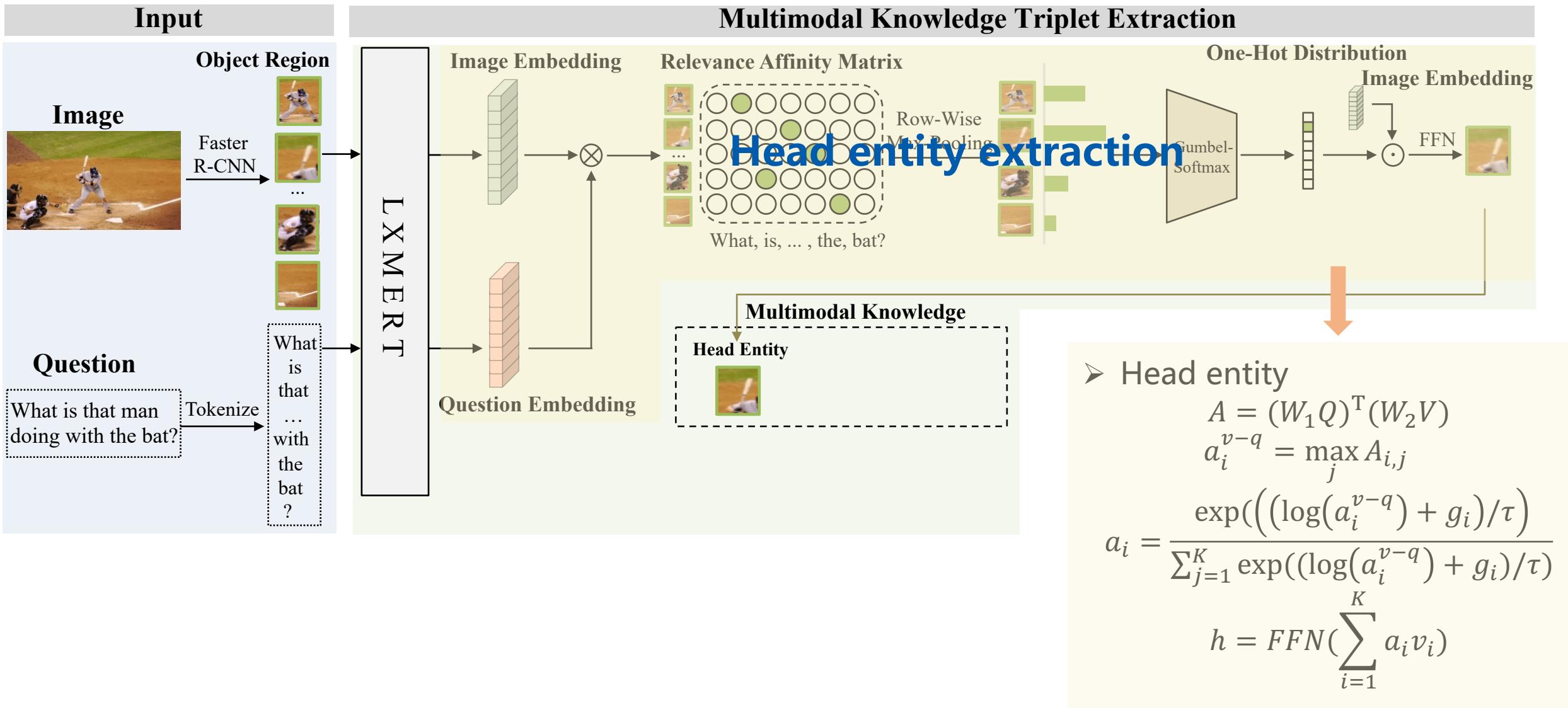
Model Framework



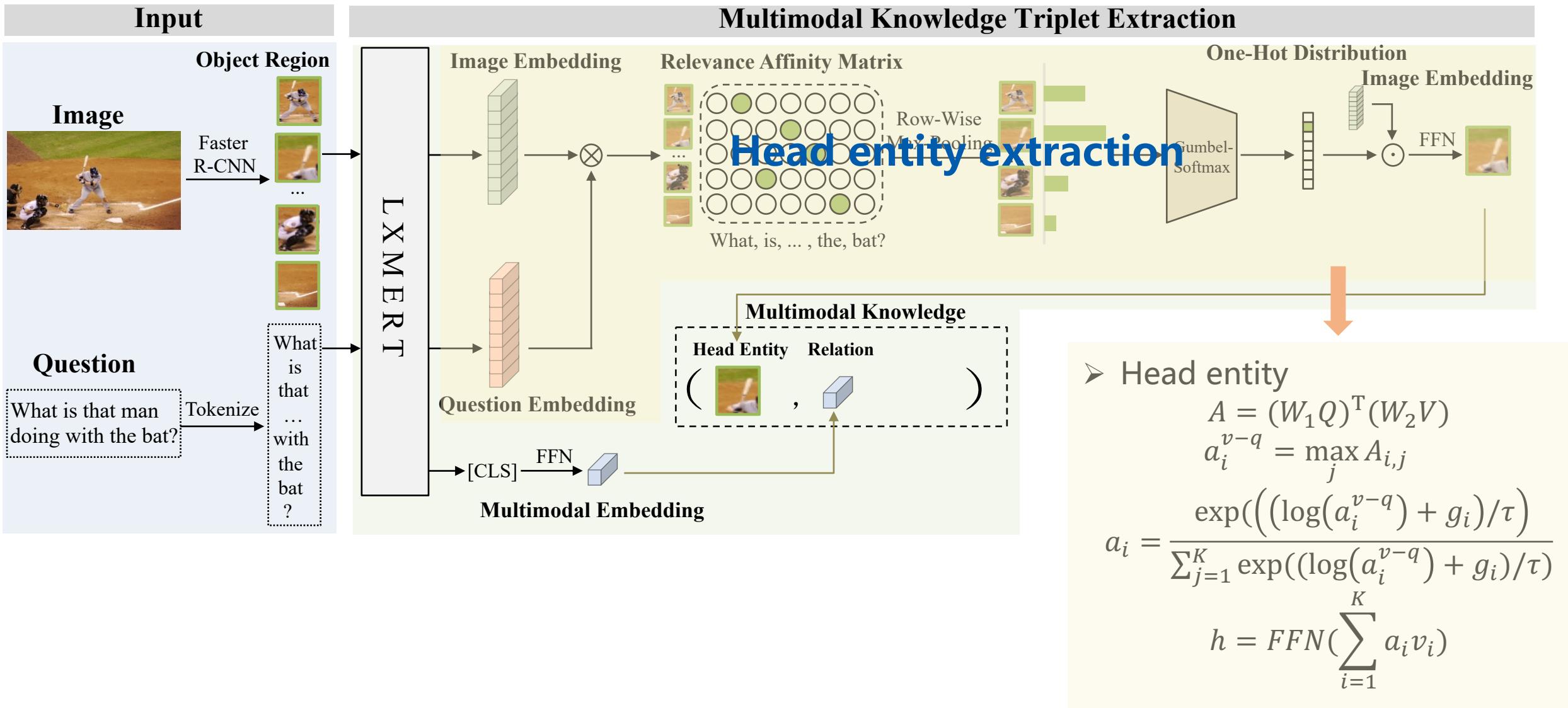
Multimodal Knowledge Triplet Extraction



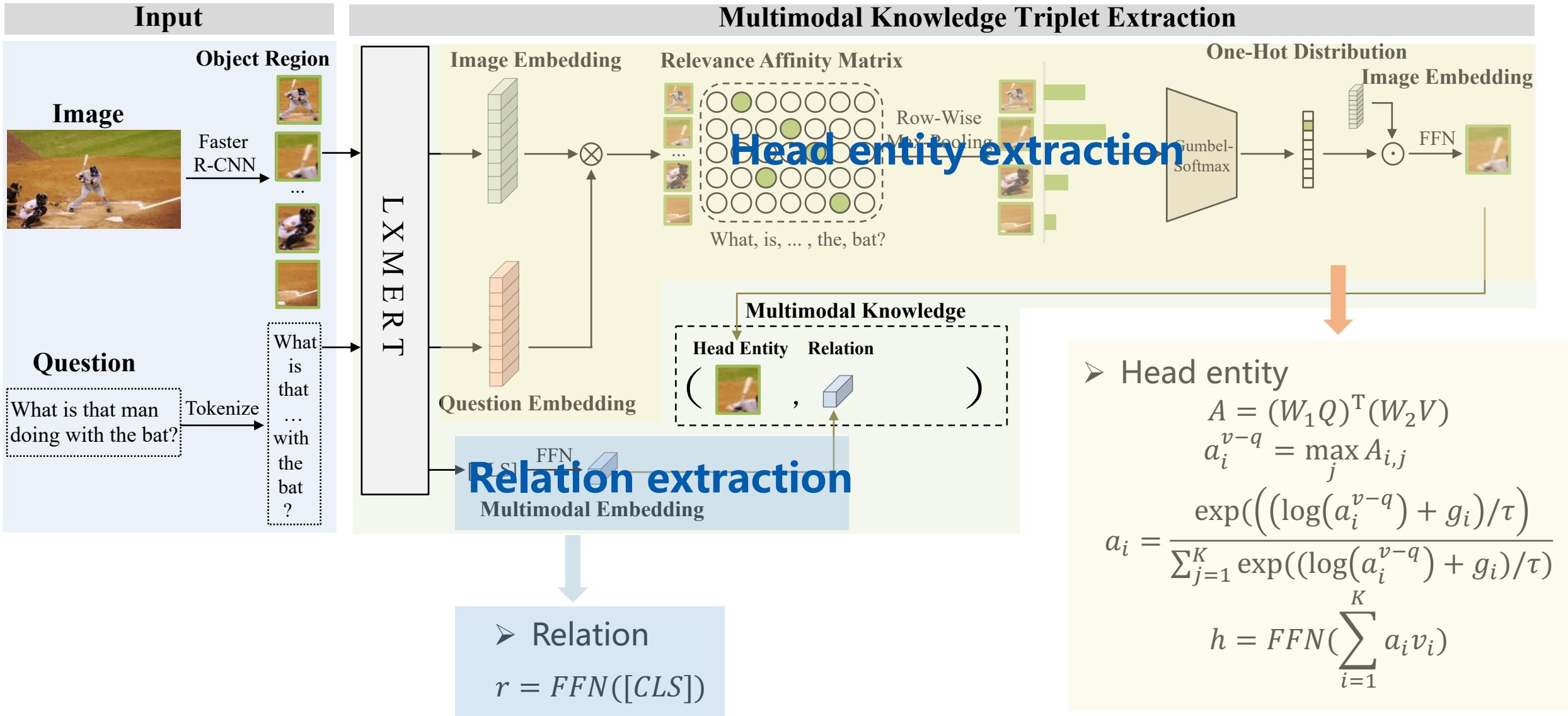
Multimodal Knowledge Triplet Extraction



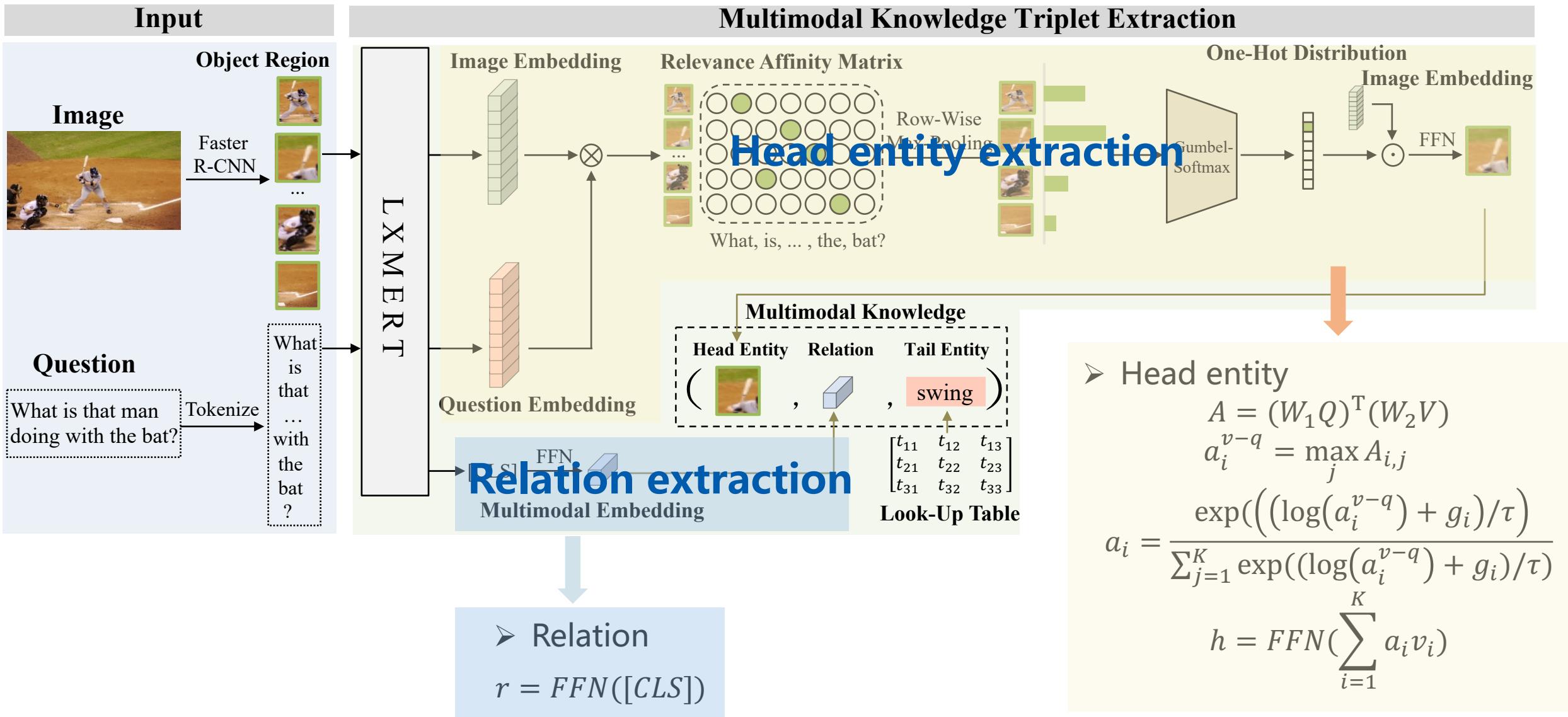
Multimodal Knowledge Triplet Extraction



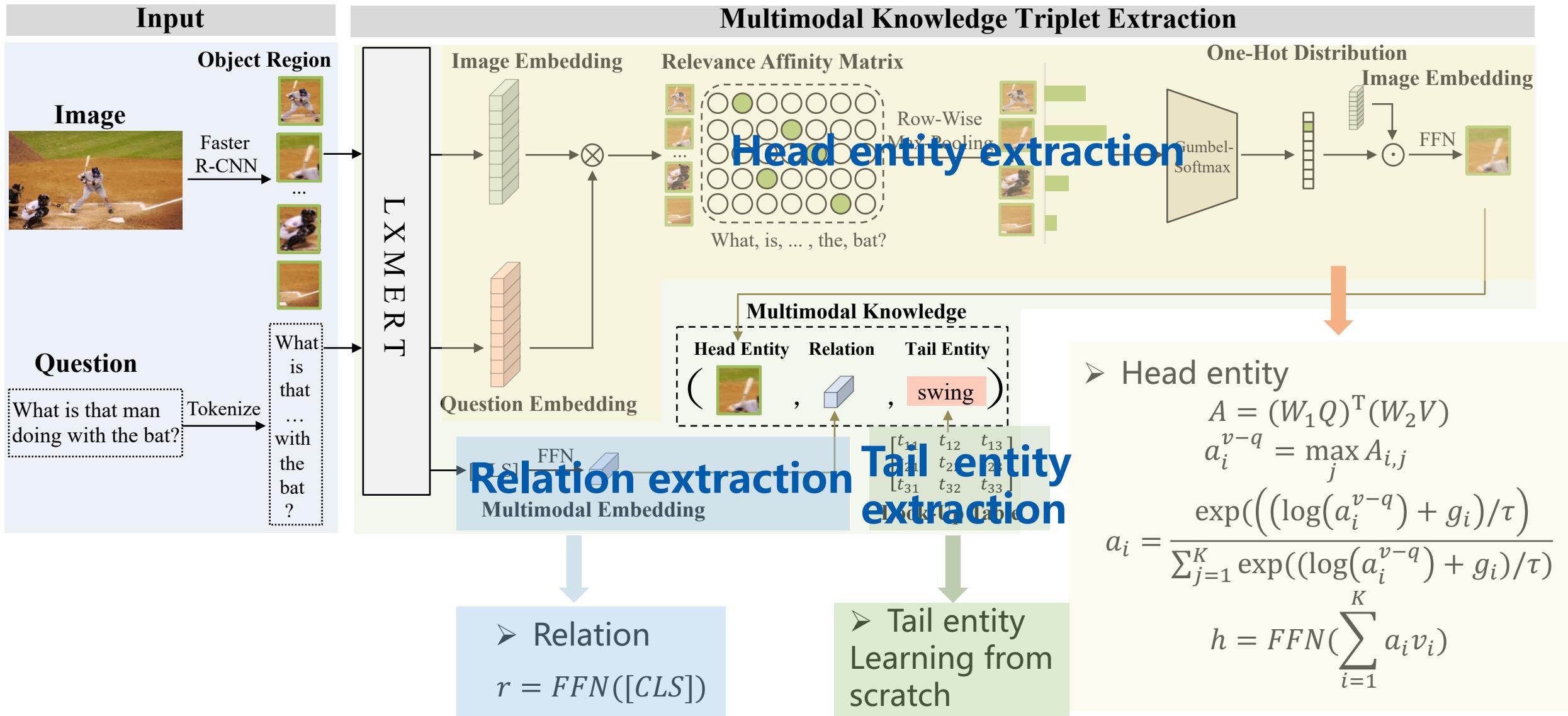
Multimodal Knowledge Triplet Extraction



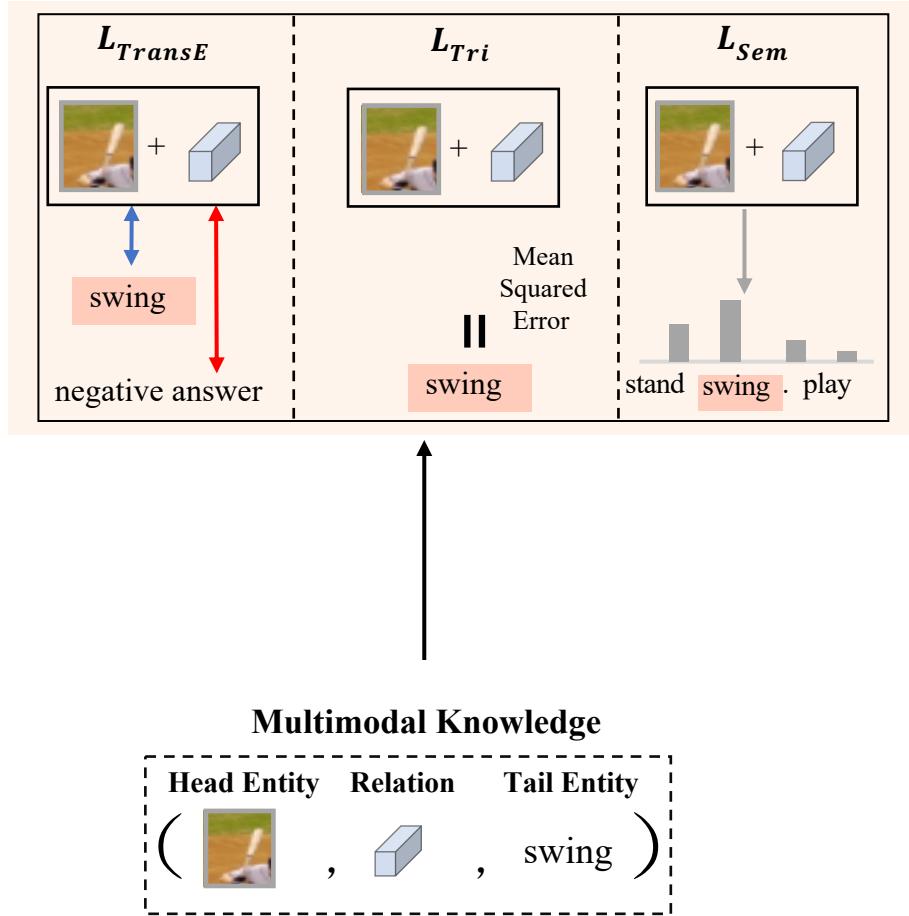
Multimodal Knowledge Triplet Extraction



Multimodal Knowledge Triplet Extraction



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:

$$P(t^+) = softmax((T)^T(h + r))$$

$$L_{Sem} = -\log(P(t^+))$$

- The final loss:

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

Knowledge Accumulation and Prediction

● Pre-training

VQA 2.0: basic visual dominant knowledge.

What is the dog wearing?
life jacket
collar



What is the girl reaching into?
bucket
apples



● Fine-tuning

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

● Inference

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

Vehicles and Transportation



Q: What sort of vehicle uses this item?
A: firetruck

Brands, Companies and Products



Q: When was the soft drink company shown first created?
A: 1898

Objects, Material and Clothing



Q: What is the material used to make the vessels in this picture?
A: copper

Sports and Recreation



Q: What is the sports position of the man in the orange shirt?
A: goalie

Cooking and Food



Q: What is the name of the object used to eat this food?
A: chopsticks

Geography, History, Language and Culture



Q: What days might I most commonly go to this building?
A: Sunday

People and Everyday Life



Q: Is this photo from the 50's or the 90's?
A: 50's

Plants and Animals



Q: What phylum does this animal belong to?
A: chordate, chordata

Science and Technology



Q: How many chromosomes do these creatures have?
A: 23

Weather and Climate



Q: What is the warmest outdoor temperature at which this kind of weather can happen?
A: 32 degrees

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}_{Tri}	41.35
3. w/o \mathcal{L}_{Sem}	42.06
4. w/o $\mathcal{L}_{\text{Tri}} \& \mathcal{L}_{\text{Sem}}$	40.84
5. w/o $\mathcal{L}_{\text{TransE}}$	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)

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

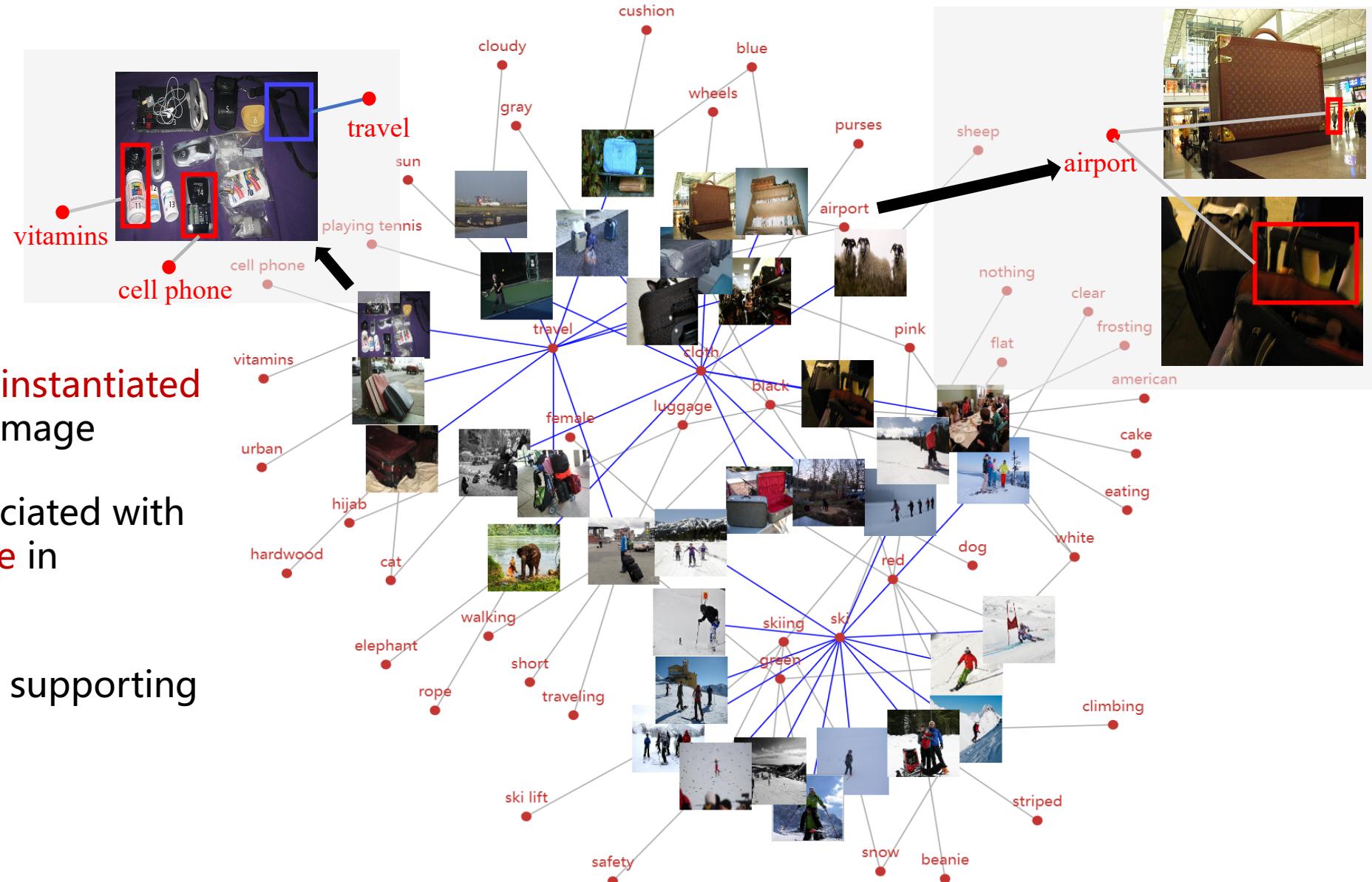
Method	Accuracy
MuKEA	42.59
MUTAN + AN*	25.43
MuKEA + (MUTAN + AN*)	35.39
MuKEA + (MUTAN + AN*) oracle	43.64
Mucko*	27.17
MuKEA + Mucko*	35.97
MuKEA + Mucko* oracle	44.84
KRISP*	32.02
MuKEA + KRISP*	37.75
MuKEA + KRISP* oracle	47.15

- **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.



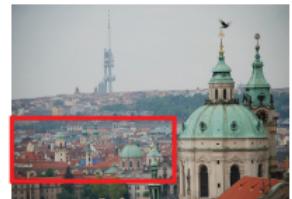
Experiment Analysis

The Predicted Multimodal Knowledge Triplets



Q: What electronic device is being featured in this photo?

KRISP: laptop ✗	MuKEA: remote ✓
Knowledge graph	Multimodal knowledge
(screen, is on, laptop) (laptop, has, screen)	(button,  , remote)
	Q: What device is pictured? Ground Truth: remote



Q: What type of architecture is shown in these buildings?

KRISP : victorian ✗	MuKEA: gothic ✓
Knowledge graph	Multimodal knowledge
(victorian, is a, comic)	(city,  , gothic)
	Q: What style of architecture is pictured in this photo? Ground Truth: gothic



Q: What style of oranges are in the stack?

KRISP : granny smith ✗	MuKEA: navel ✓
Knowledge graph	Multimodal knowledge
(apple, capable of, granny smith)	(orange ,  , navel)
	Q: What kind of orange is this? Ground Truth: navel



Q: What kind of plane is this?

KRISP: biplane ✗	MuKEA: prop plane ✓
Knowledge graph	Multimodal knowledge
(biplane, is a, airplane)	(propeller,  , prop plane)
	Q: What type of fuel does this plane use? Ground Truth: jet



Q: Why is this dangerous?

KRISP : danger ✗	MuKEA: drown ✓
Knowledge graph	Multimodal knowledge
(danger, has property, bad)	(water,  , drown)
	Q: What is the largest one of these natural occurrences ever recorded? Ground Truth: 100 feet



Q: What is the name for a child of the species shown?

KRISP : herd ✗	MuKEA: calf ✓
Knowledge graph	Multimodal knowledge
(sheep, is in, herd) (herd, has part, lamb)	(cow,  , calf)
	Q: The baby of this animal is called what? Ground Truth: calf

- MuKEA captures instantiated knowledge

- MuKEA contains multi-object involved complex knowledge

- MuKEA avoids the cascading error.

Experiment Analysis

Progressive Knowledge Accumulation

VQA 2.0 samples



Q: What kind of vehicle is shown?
A: motorcycle



Q: What is the container made of?
A: plastic

Knowledge after pre-training



what motorcycle

OK-VQA samples

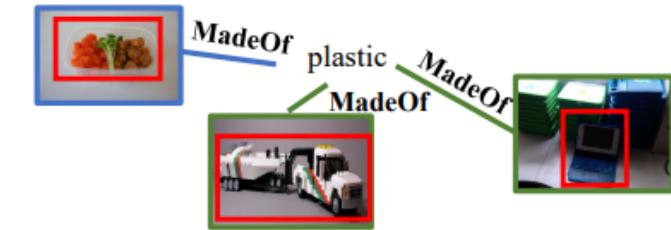
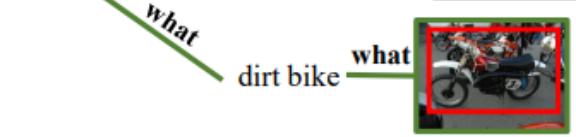
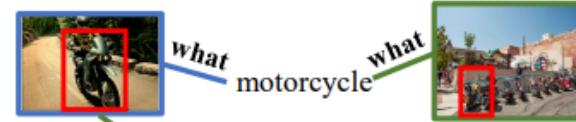


Q: What are parked in the street?
A: motorcycle



Q: What material is the screen depicted made of?
A: plastic

Knowledge after fine-tuning



- 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

VQA 2.0 samples



Q: What type of animal is in the picture?
A: giraffe



Q: Where is the train?
A: on tracks

OK-VQA samples

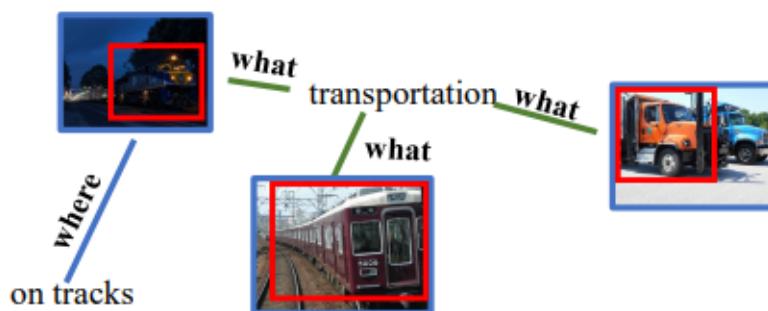
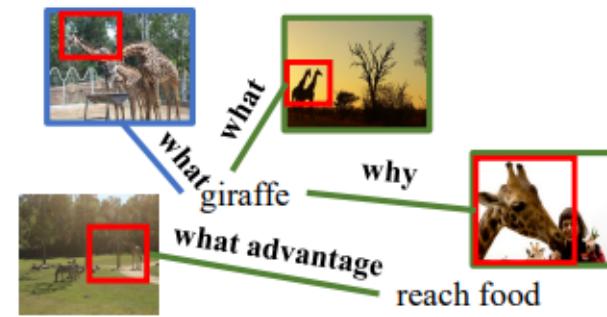


Q: What evolutionary advantage does the neck of a giraffe give it?
A: reach food



Q: What kind of train is this?
A: transportation

Knowledge after fine-tuning



Test samples



Q: Which animal in the picture has a neck that evolved to reach food?
MuKEA: giraffe ✓



Q: What is the function of the object on tracks?
MuKEA: transportation ✓

- 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/>



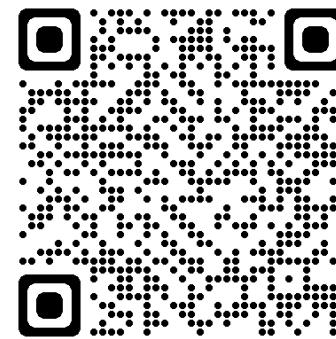
中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS



Homepage



Paper



Code

