



# 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

Université   
de Montréal



THE UNIVERSITY  
of ADELAIDE

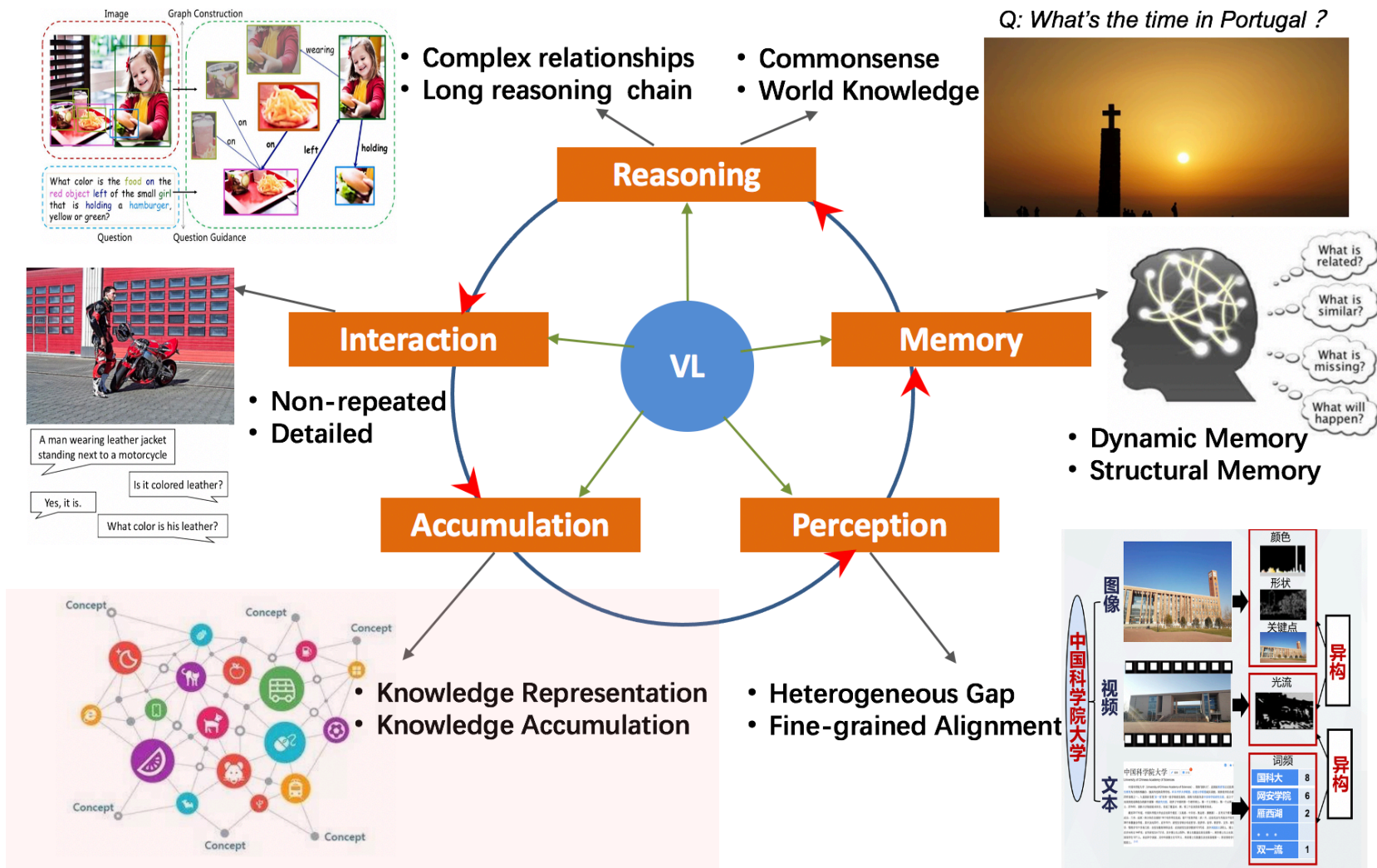


# Content

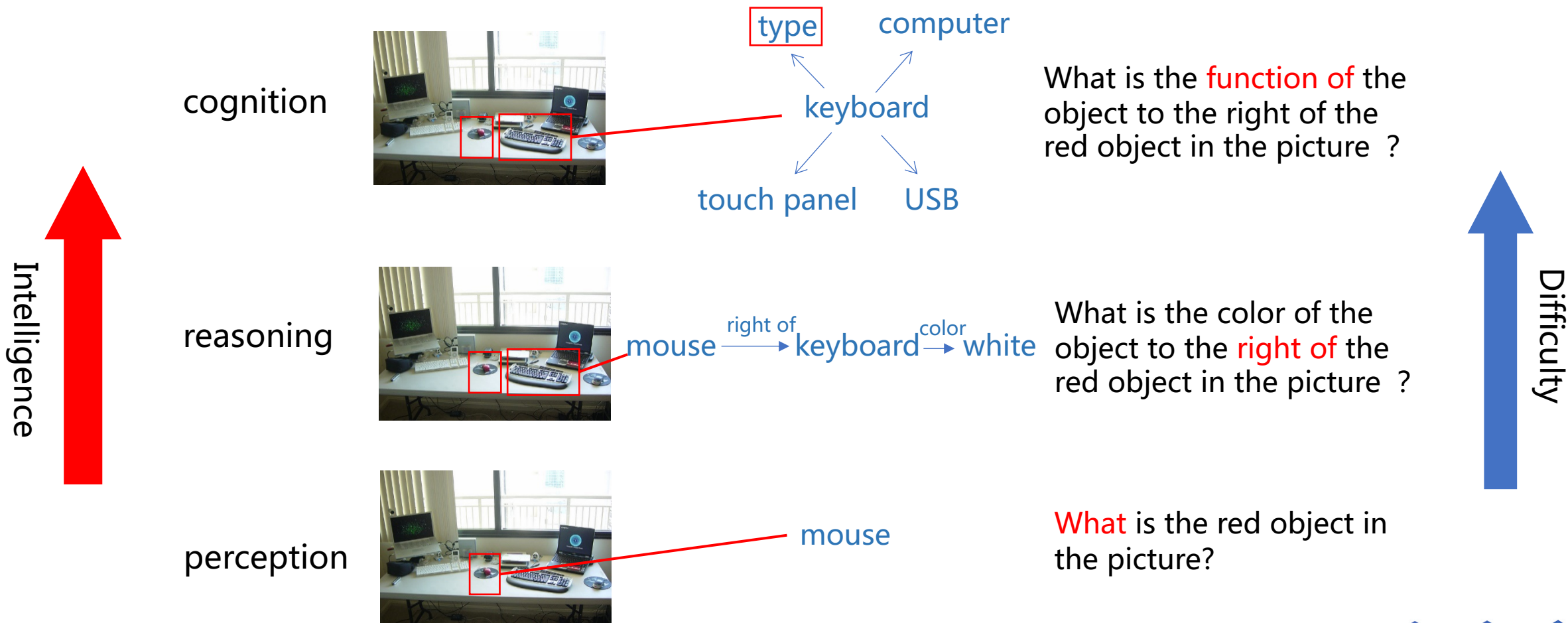
- Motivation
- Model
- Experiments
- Summary



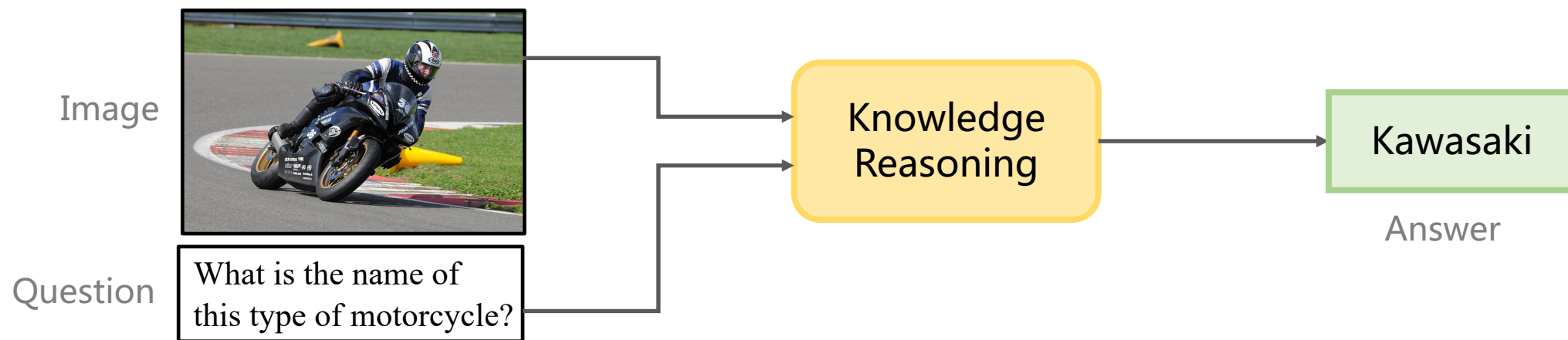
# ◆ The roadmap of our CogModal group



- ◆ 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**

WIKIPEDIA The Free Encyclopedia

Google

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

ConceptNet 5  
Computer and get connect

DBpedia Indonesian

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

GPT-3

Translate English to German: "hat so good."

"hat sentence: The cover is being well."

"hat sentence: The film is good on the grass, instead, it stays in grassy in a field."

T5

"hat sentence: state authorities dispatched emergency crews Tuesday to survey the damage after an onslaught of severe weather in Mississippi."

"hat sentence: get go."

"hat sentence: not acceptable."

T5

"hat sentence: hospital after a storm in stable crafty."

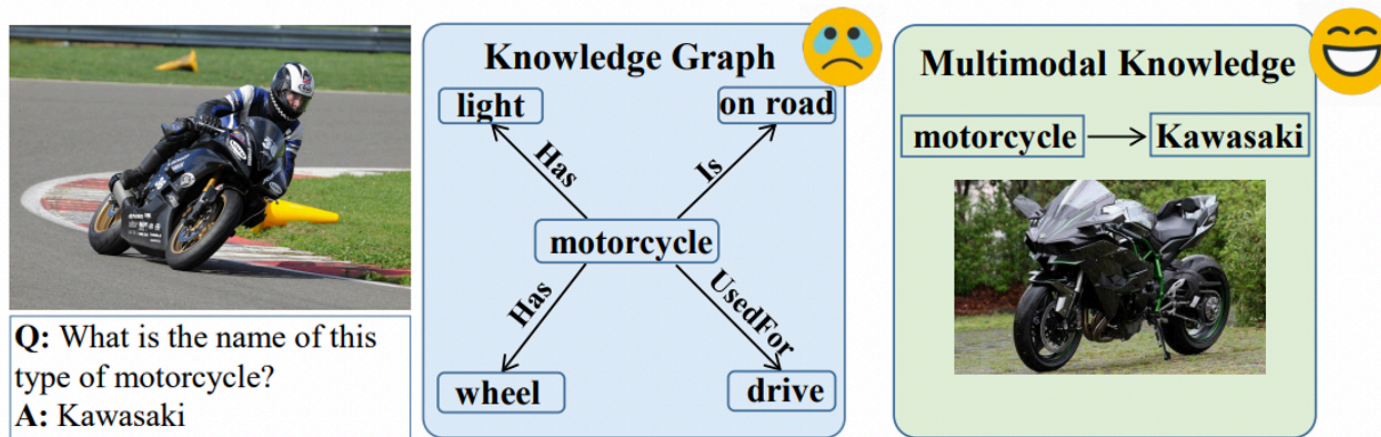
- PICa [EMNLP2022]
- Frozen [NIPS2021]
- KAT [arXiv2022]

motorcycle → Kawasaki

motorcycle image with a red question mark.

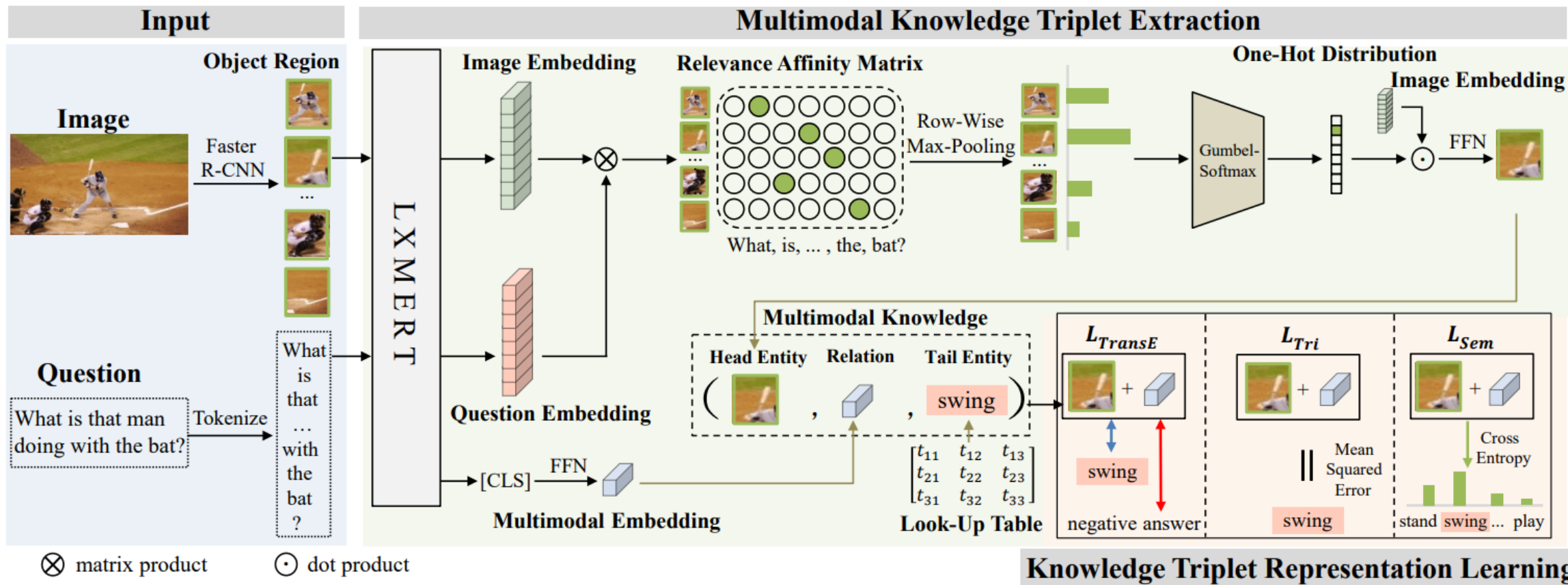
- KM<sup>4</sup> [Information fusion2021]
- Gaia [ACL2020]
- MKGAT [CIKM2020]

# Our Goal

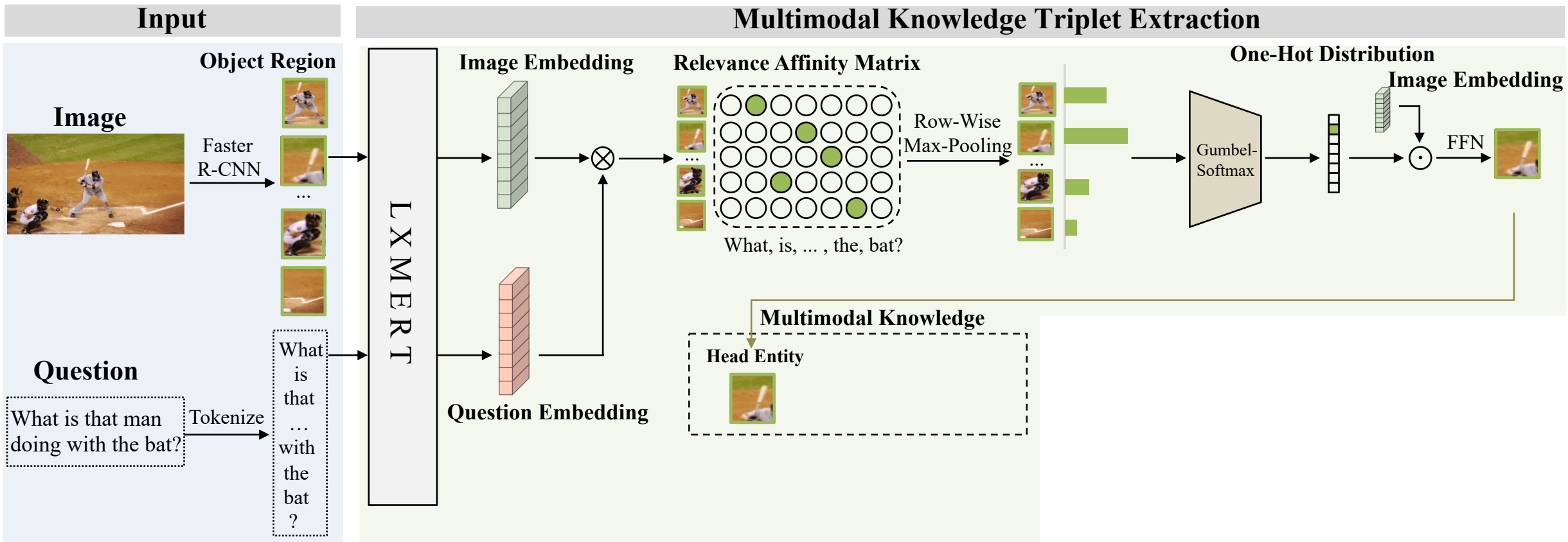


- How to **represent** the multimodal knowledge?
- How to **accumulate** the multimodal knowledge in the VQA scenarios?
- How to maintain the advantages of traditional knowledge graph in **explainable reasoning**?

# Model Framework

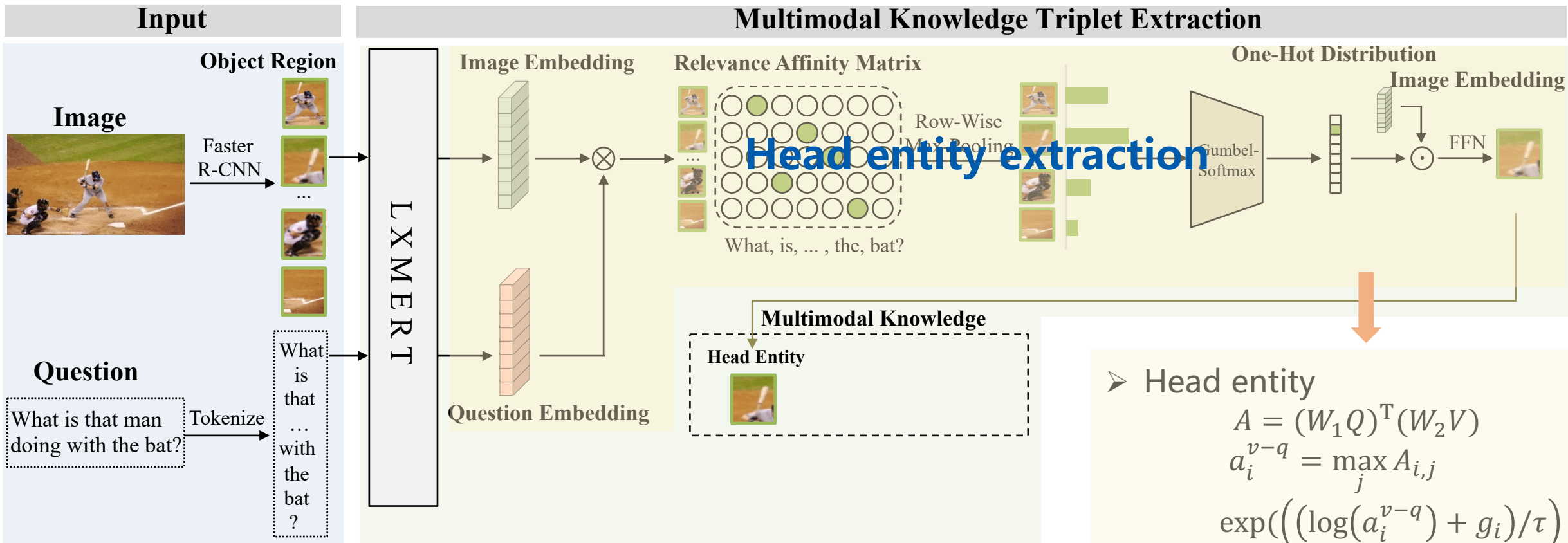


# Multimodal Knowledge Triplet Extraction





# Multimodal Knowledge Triplet Extraction



➤ Head entity

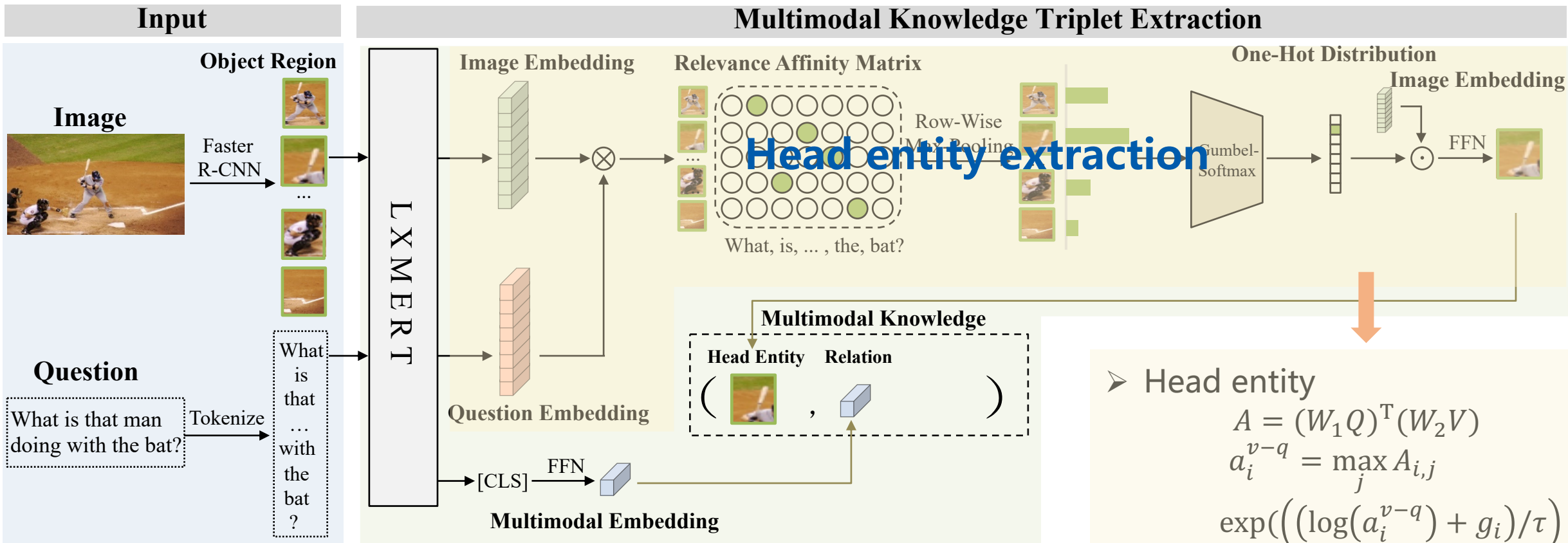
$$A = (W_1 Q)^T (W_2 V)$$

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

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

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

# Multimodal Knowledge Triplet Extraction



➤ Head entity

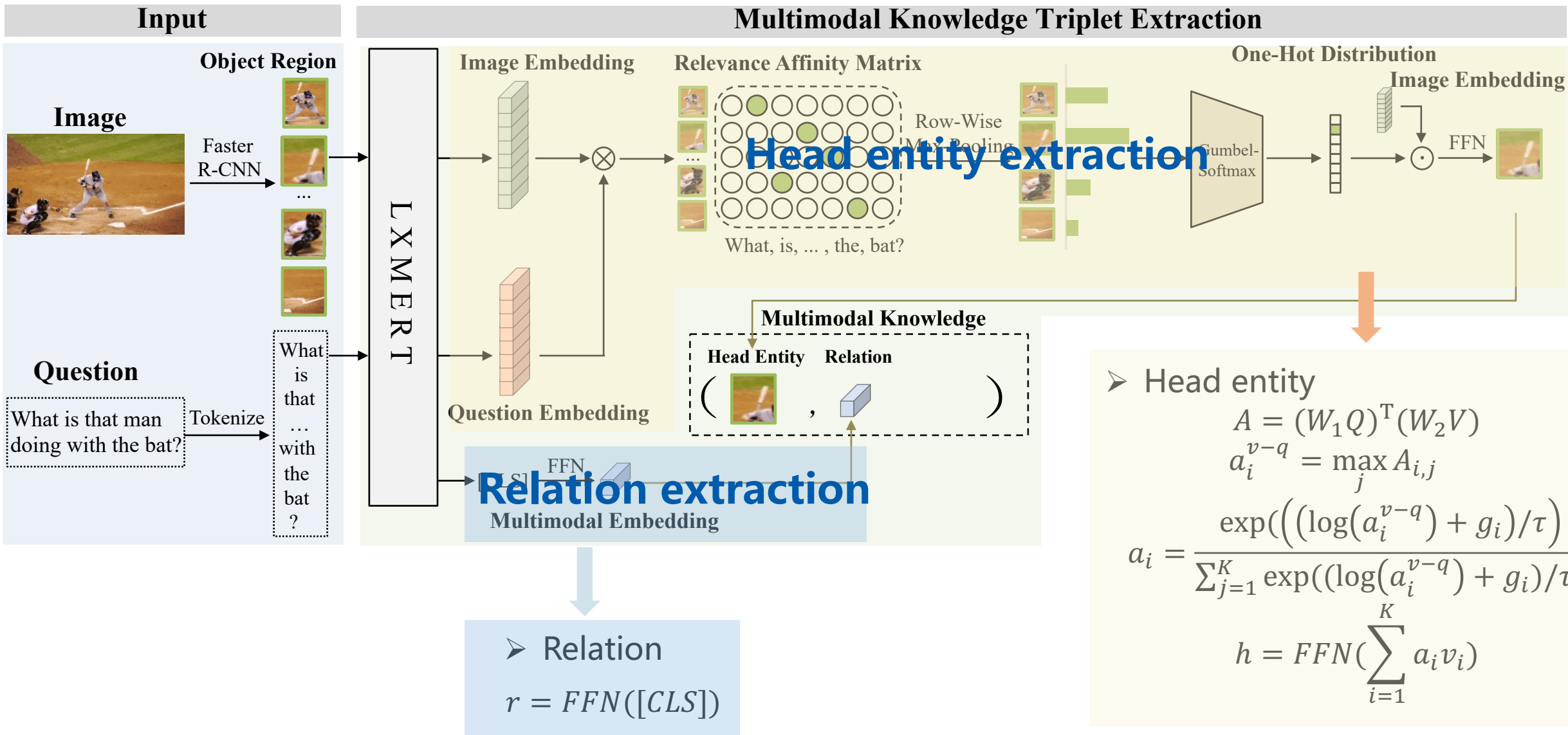
$$A = (W_1 Q)^T (W_2 V)$$

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

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

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

# Multimodal Knowledge Triplet Extraction



➤ Head entity

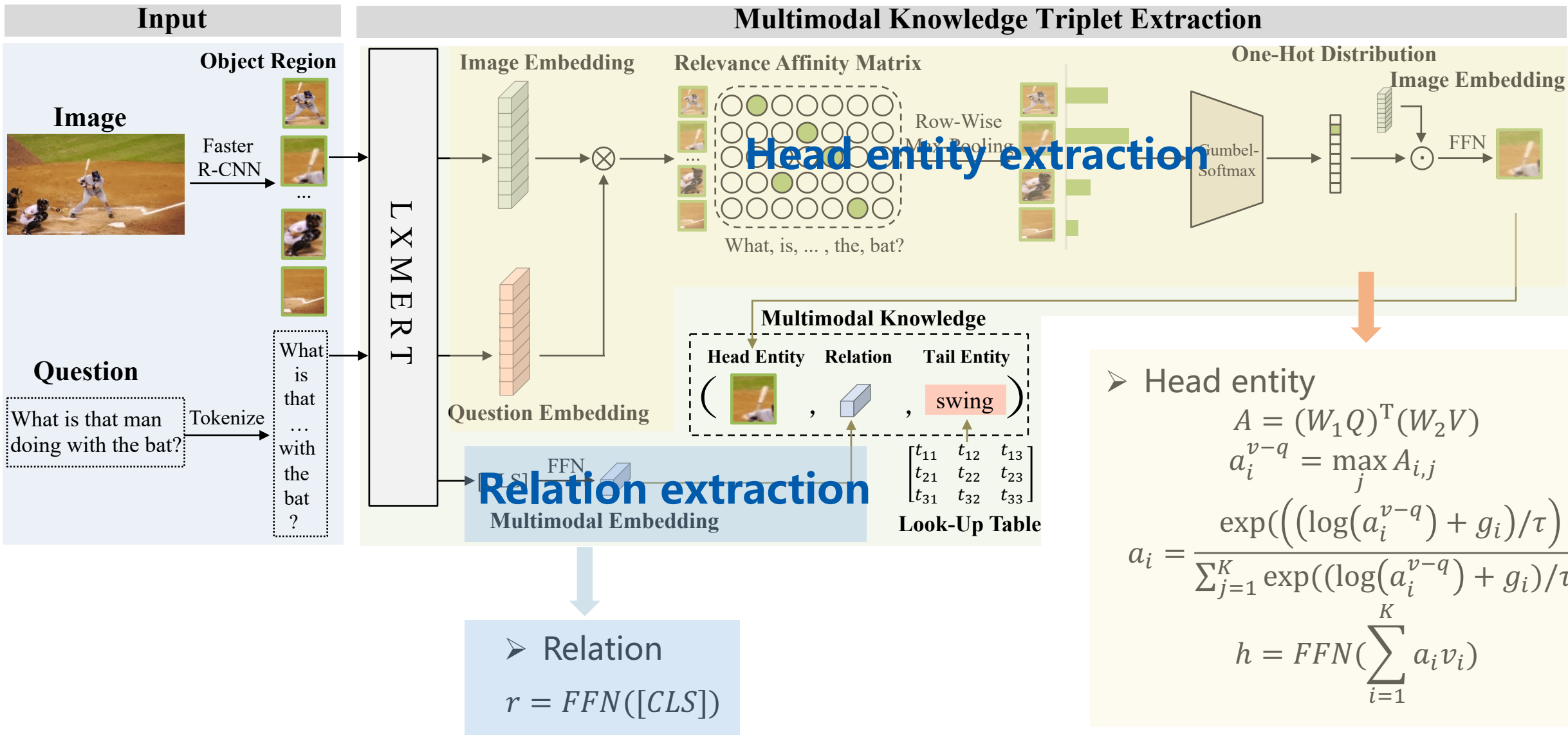
$$A = (W_1 Q)^T (W_2 V)$$

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

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

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

# Multimodal Knowledge Triplet Extraction



➤ Head entity

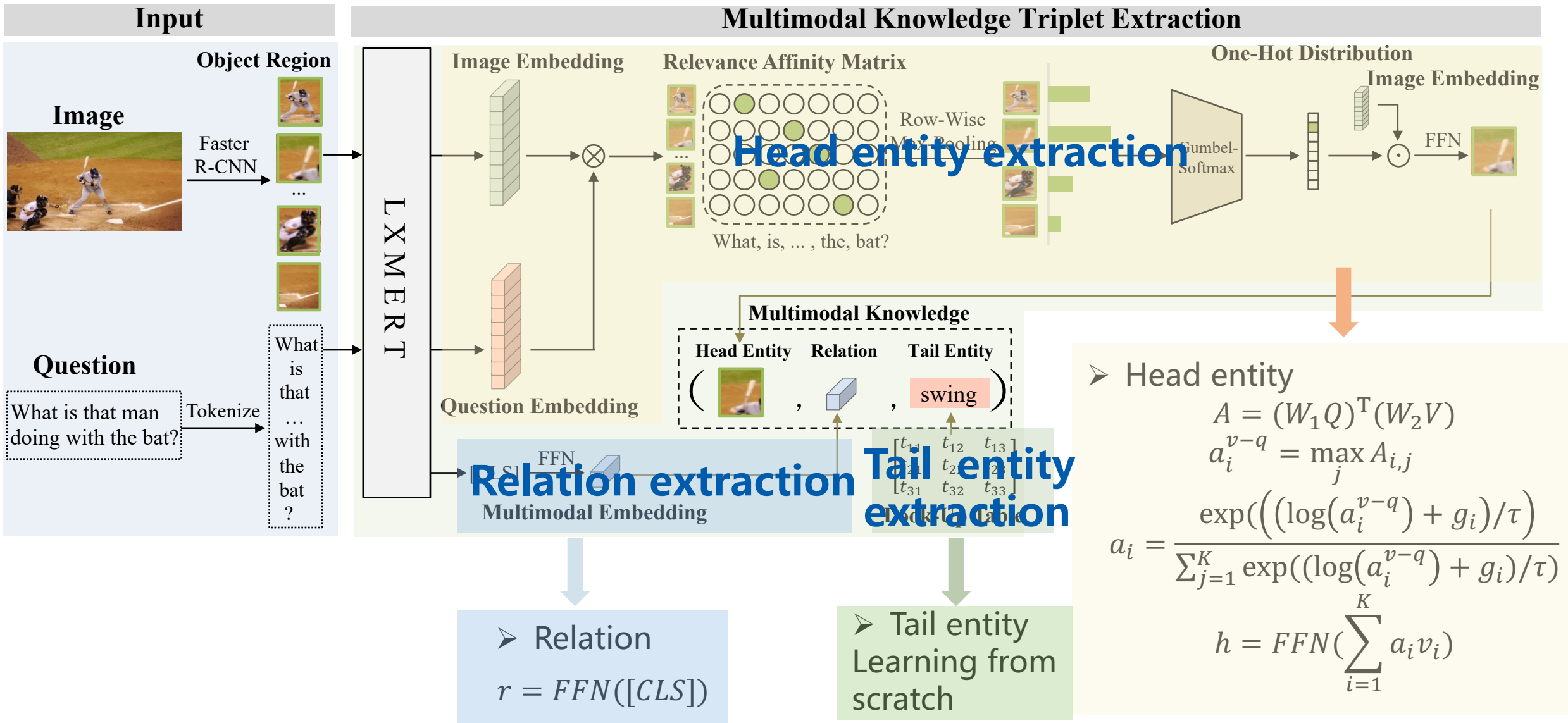
$$A = (W_1 Q)^T (W_2 V)$$

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

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

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

# Multimodal Knowledge Triplet Extraction



➤ Head entity

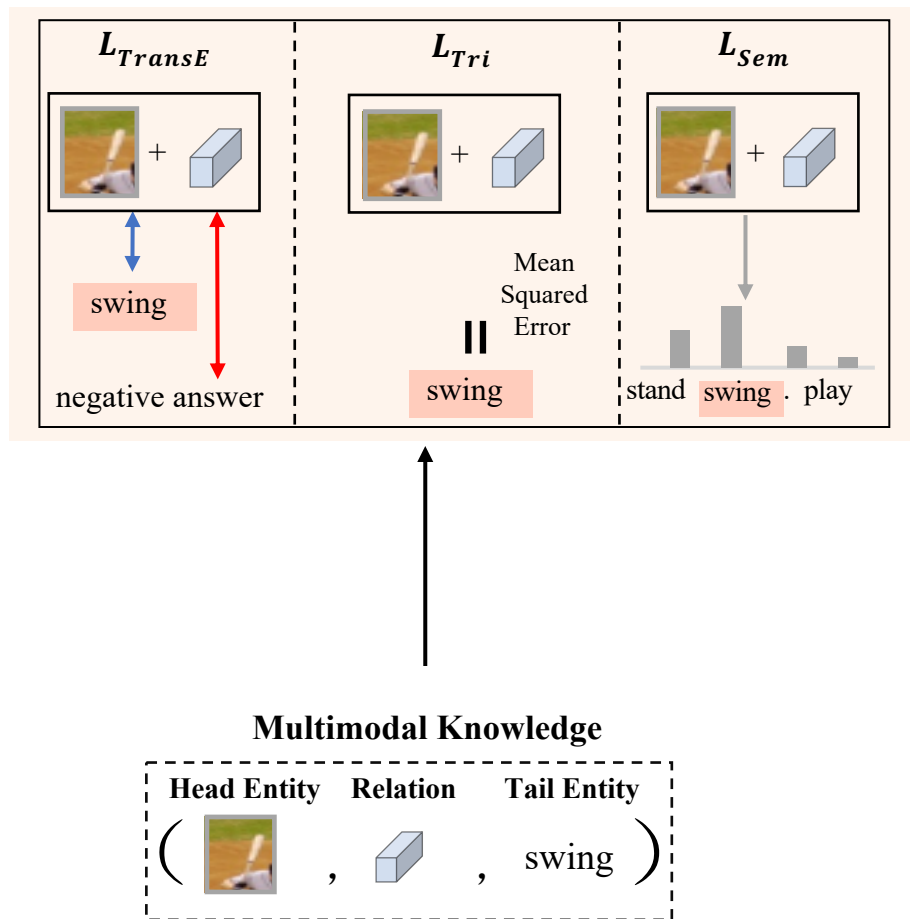
$$A = (W_1 Q)^T (W_2 V)$$

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

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

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

# 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:

$$P(t^+) = \text{softmax}((T)^T (h + r))$$

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

- The final loss:

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

# Knowledge Accumulation and Prediction

- **Pre-training**

VQA 2.0: basic visual dominant knowledge.

- **Fine-tuning**

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

- **Inference**

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



<p>Vehicles and Transportation</p> <p>Q: What sort of vehicle uses this item? A: firetruck</p>	<p>Brands, Companies and Products</p> <p>Q: When was the soft drink company shown first created? A: 1898</p>	<p>Objects, Material and Clothing</p> <p>Q: What is the material used to make the vessels in this picture? A: copper</p>	<p>Sports and Recreation</p> <p>Q: What is the sports position of the man in the orange shirt? A: goalie</p>	<p>Cooking and Food</p> <p>Q: What is the name of the object used to eat this food? A: chopsticks</p>
<p>Geography, History, Language and Culture</p> <p>Q: What days might I most commonly go to this building? A: Sunday</p>	<p>People and Everyday Life</p> <p>Q: Is this photo from the 50's or the 90's? A: 50's</p>	<p>Plants and Animals</p> <p>Q: What phylum does this animal belong to? A: chordate, chordata</p>	<p>Science and Technology</p> <p>Q: How many chromosomes do these creatures have? A: 23</p>	<p>Weather and Climate</p> <p>Q: What is the warmest outdoor temperature at which this kind of weather can happen? A: 32 degrees</p>

# 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 <sup>4</sup> [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	<b>42.59</b>

- 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	<b>9.84</b>	9.24	5.51	13.89	21.30
<b>MuKEA</b>	<b>59.12</b>	<b>44.88</b>	<b>37.36</b>	<b>52.47</b>	<b>48.08</b>	<b>35.63</b>	<b>31.61</b>	<b>17.62</b>	6.14	<b>9.85</b>	<b>6.22</b>	<b>18.28</b>	<b>27.38</b>

- 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)	<b>42.59</b>
<b>Ablation of Loss Function</b>	
2. w/o $\mathcal{L}_{\text{Tri}}$	41.35
3. w/o $\mathcal{L}_{\text{Sem}}$	42.06
4. w/o $\mathcal{L}_{\text{Tri}}$ & $\mathcal{L}_{\text{Sem}}$	40.84
5. w/o $\mathcal{L}_{\text{TransE}}$	24.50
<b>Ablation of Triplet Representation</b>	
6. head entity w/ soft-attention	40.67
7. relation w/ self-attention	40.79
8. tail entity w/ GloVe	41.42
<b>Ablation of Triplet Structure</b>	
9. w/o $h$	39.83
10. w/o $r$	39.40
<b>Ablation of Knowledge Source</b>	
11. w/o VQA 2.0 knowledge	36.35
12. w/o OK-VQA knowledge	27.20
<b>Ablation of Pre-training Knowledge</b>	
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

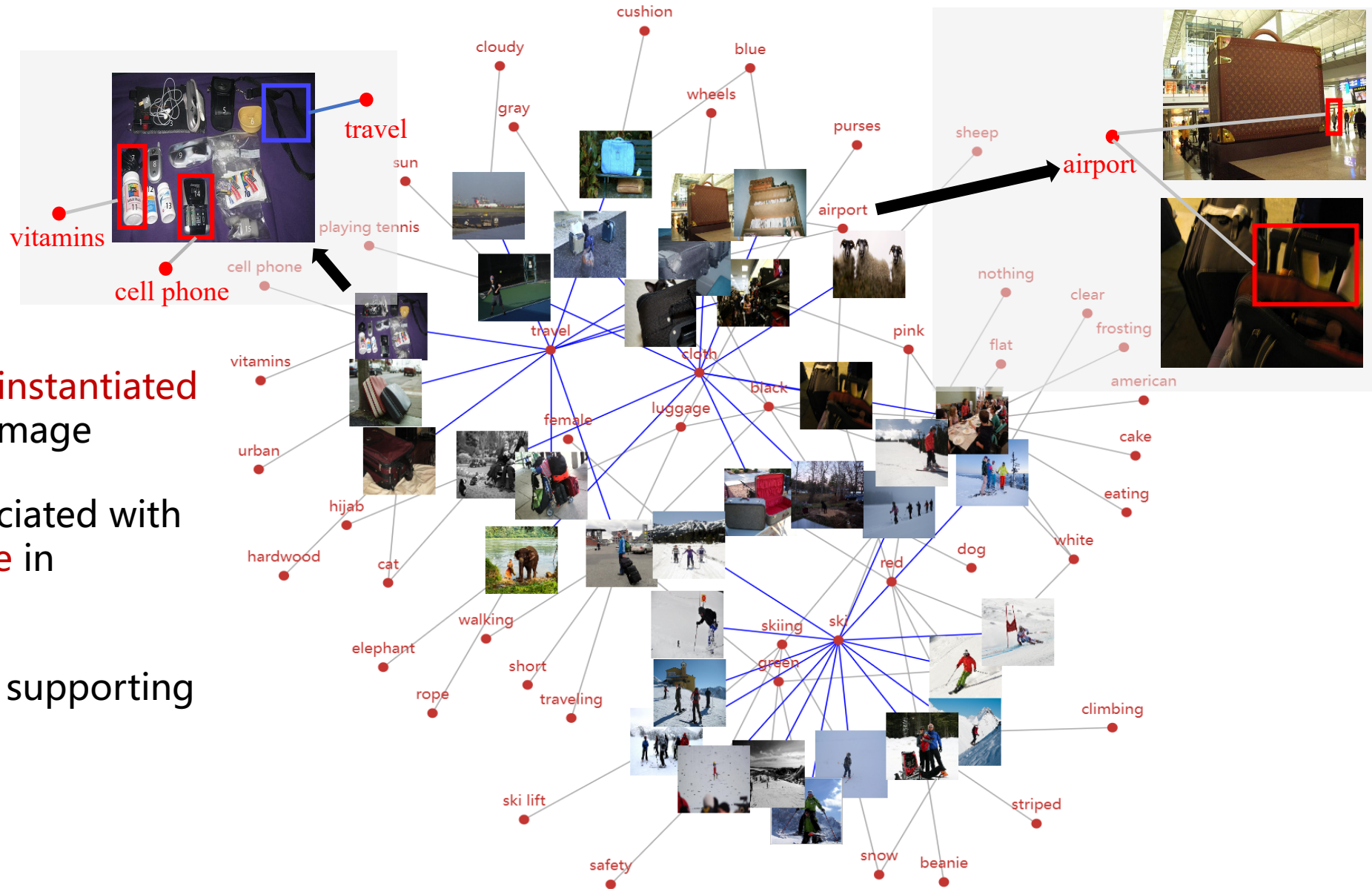
Method	Accuracy
MuKEA	42.59
MUTAN + AN*	25.43
MuKEA + (MUTAN + AN*)	35.39
MuKEA + (MUTAN + AN*) oracle	43.64
Mucko*	27.17
MuKEA + Mucko*	35.97
MuKEA + Mucko* oracle	44.84
KRISP*	32.02
MuKEA + KRISP*	37.75
MuKEA + KRISP* oracle	47.15

- **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





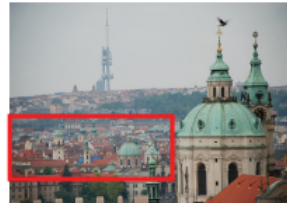
Q: What electronic device is being featured in this photo?

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





Q: What kind of plane is this?

<b>KRISP:</b> biplane ✗	<b>MuKEA:</b> 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: What type of architecture is shown in these buildings?

<b>KRISP :</b> victorian ✗	<b>MuKEA:</b> 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: Why is this dangerous?

<b>KRISP :</b> danger ✗	<b>MuKEA:</b> 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 style of oranges are in the stack?

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



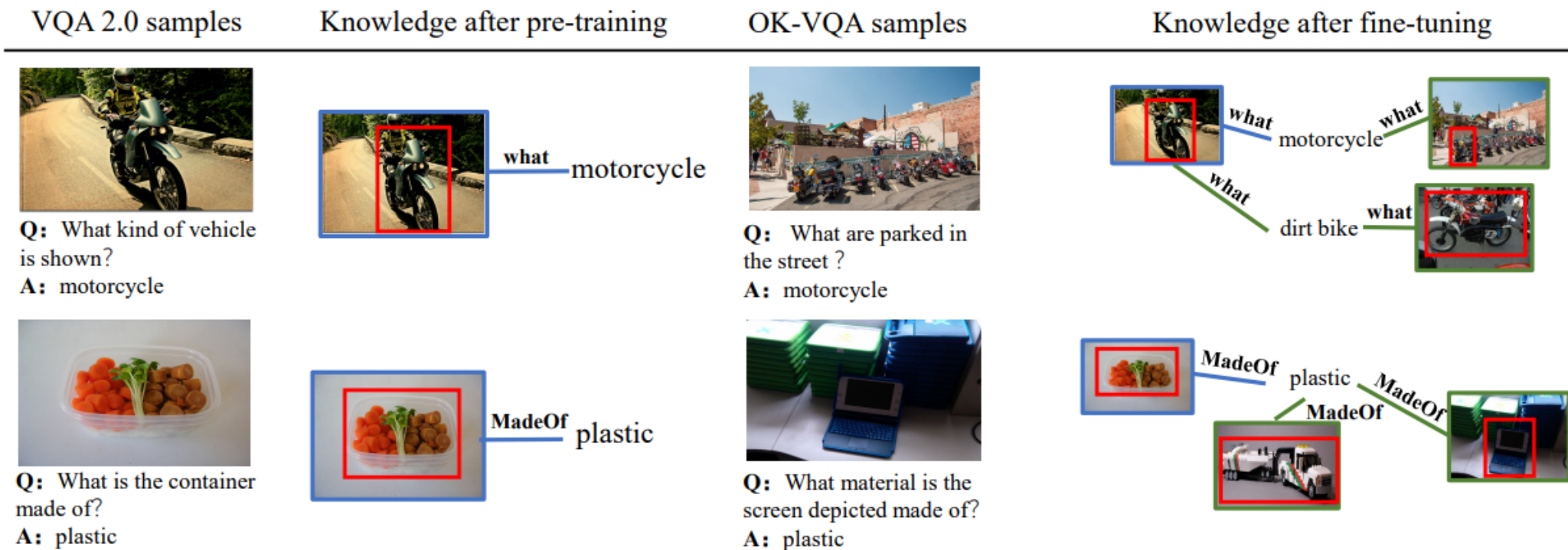
Q: What is the name for a child of the species shown?

<b>KRISP :</b> herd ✗	<b>MuKEA:</b> 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







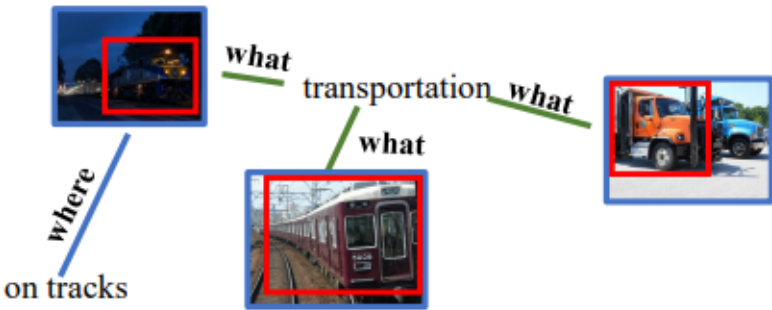

## Progressive Knowledge Accumulation



- 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	OK-VQA samples	Knowledge after fine-tuning	Test samples
 <p><b>Q:</b> What type of animal is in the picture? <b>A:</b> giraffe</p>	 <p><b>Q:</b> What evolutionary advantage does the neck of a giraffe give it? <b>A:</b> reach food</p>		 <p><b>Q:</b> Which animal in the picture has a neck that evolved to reach food? <b>MuKEA:</b> giraffe ✓</p>
 <p><b>Q:</b> Where is the train? <b>A:</b> on tracks</p>	 <p><b>Q:</b> What kind of train is this? <b>A:</b> transportation</p>		 <p><b>Q:</b> What is the function of the object on tracks? <b>MuKEA:</b> transportation ✓</p>

- 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](mailto: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

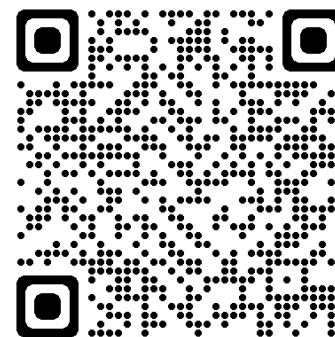


中国科学院大学  
University of Chinese Academy of Sciences

Homepage



Paper



Code

