Multimodal Knowledge Graphs:

Construction, Inference, and Challenges

Meng Wang

Guilin Qi, Qiushuo Zheng, Chaoyu Bai

Southeast University

- Multimodality
- Multimodal KG Construction
- Inference
- Challenges

What is Multimodal Knowledge?

Multimodality: is the application of multiple literacies within one medium[1].

Wikipedia
The Free Encyclopedia

Knowledge: Facts acquired through experience or education; the theoretical or practical understanding of a subject

------Oxford dictionary (English) , 2016

[1] https://en.wikipedia.org/wiki/Multimodality

Multimodal knowledge:
is an awareness or understanding of someone or something in different multimodalities.

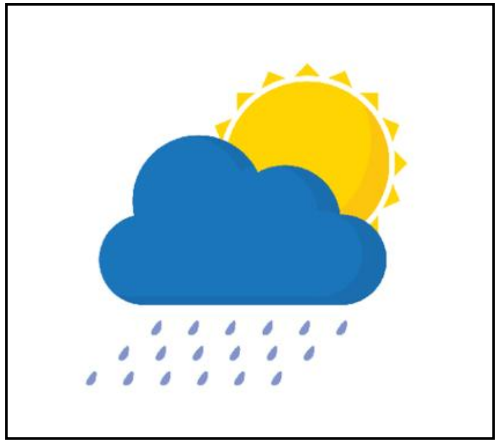
We can extract different multimodal knowledge on a same fact (or a same conventional knowledge)

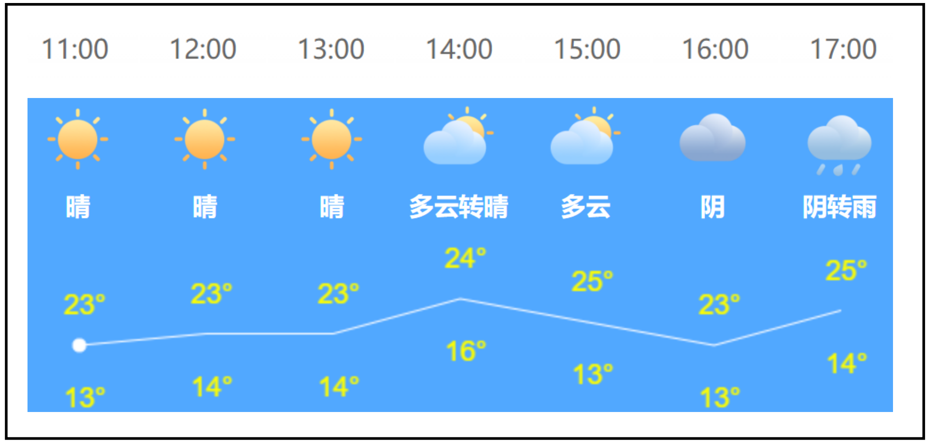
We can use different multimodal knowledge to understand on a same thing (a conventional entity)

Case 1:

南昌天气晴，局部地区有短时降雨

Spoken and Visual Knowledge

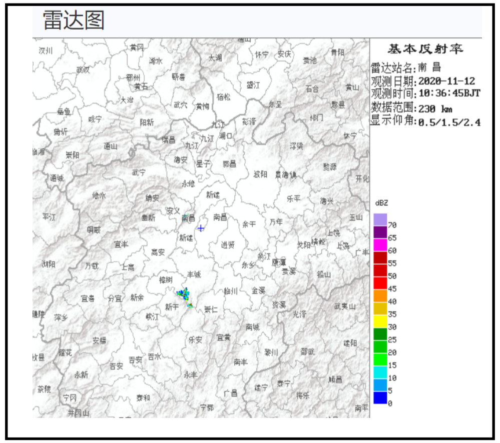




Symbol and Temperature Knowledge

欢迎大家参加CCKS，江西省最近到秋天了，所以南昌这几天都是好天！

Gestural and Linguistic Knowledge

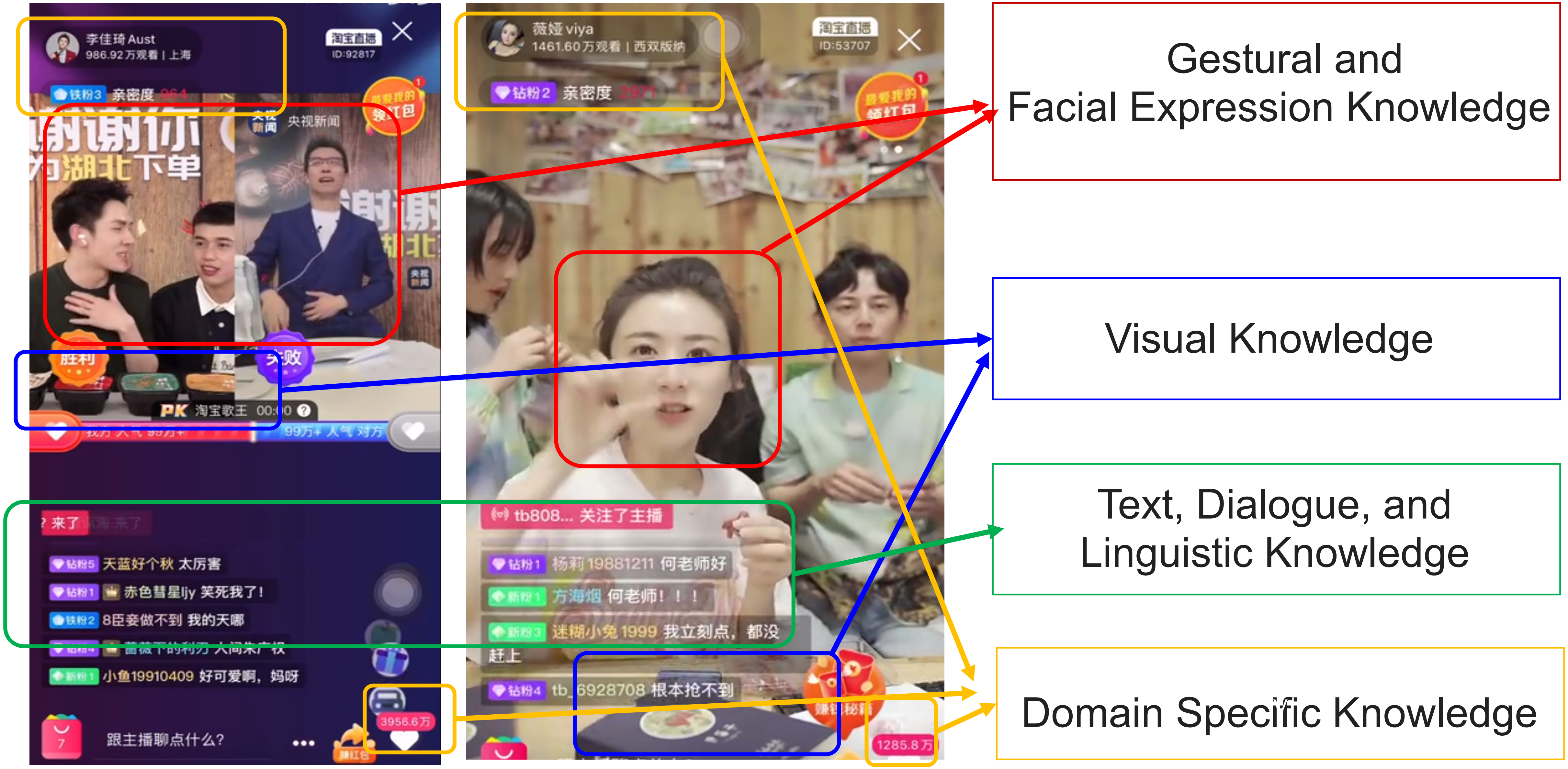


Geography Knowledge

Weather Specific Knowledge

（图片来源于网络，仅供示意）

Case 2:
这款商品真的很不错！



We can extract different multimodal knowledge on a same fact (or a same conventional knowledge)

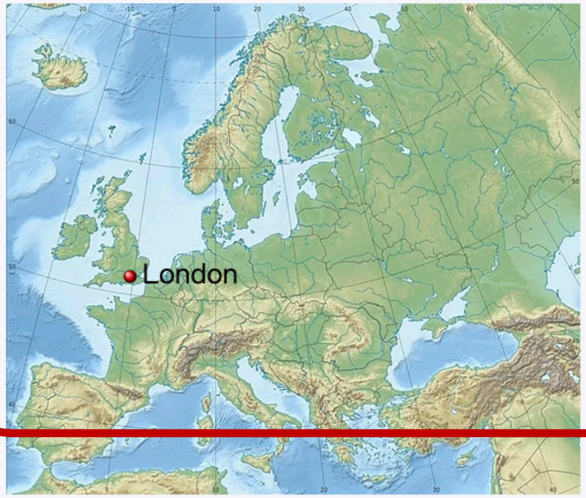
We can use different multimodal knowledge to understand on a same thing (a conventional entity)

Case 3:

London

Capital city

Clockwise from top: City of London in the foreground with Canary Wharf in the far background, Trafalgar Square, London Eye, Tower Bridge and a London Underground roundel in front of Elizabeth Tower



Location within Europe

Show map of the United Kingdom

Show map of Europe

Show map of Earth

Show all

Coordinates: 51°30′26″N 0°7′39″W

Sovereign State	United Kingdom
Constituent Country	England
Region	London (coterminous)
Counties	Greater London City of London
Settled by Romans	AD 47[1] *as Londinium*
Districts	City of London & 32 boroughs
Government	
• **Type**	Executive mayoralty and deliberative assembly within unitary constitutional monarchy

Image

Text

KG

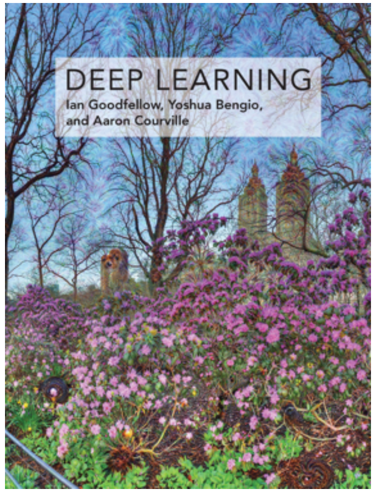
Case 4:

科比携手邓肯KG进入名人堂候选名单 三大传奇2020年将圆梦
2019
12/20
06:18
猫大熊
企鹅号
科比携手邓肯KG进入名人堂候选名单 三大传奇2020年将圆梦
自动播放
分享
评论
785
北京时间12月20日，奈史密斯篮球名人堂公布了2020年篮球名人堂候选人名单，其中包括科比、邓肯、加内特这三大NBA超级巨星，可谓是史上最星光熠熠的一届。
赛事精选
12/20 12/21 12/22
76人 0 未开始 09:00 独行侠 0
骑士 0 未开始 08:00 灰熊 0
步行者 0 未开始 08:00 国王 0
凯尔特人 0 未开始 08:30 活塞 0
猛龙 0 未开始 08:30 奇才 0
全部赛程
推荐视频
【球星】哈登集锦 53分16板17助历史第一人
【扣篮】位置感极佳 特纳底线埋伏好冲起补扣
【球星】庄神集锦 砍下25分18个篮板频频双手爆筐
【集锦】尼克斯122-129火箭 哈登生涯新高53+16+17破三分纪录
1日NBA头条 哈登53+16+17历史第一 韦少半场三双

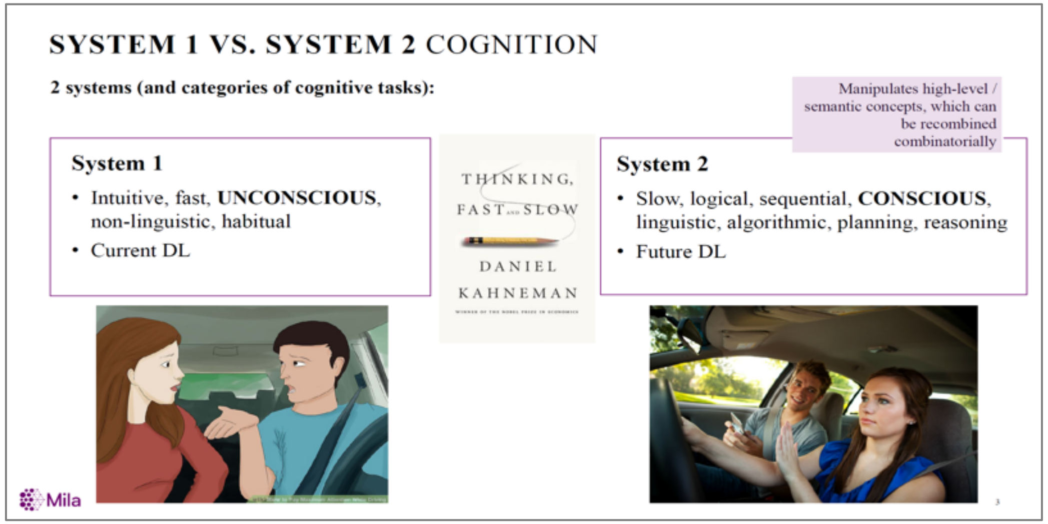
圣诞节大型立体粉色豪华圣诞树家用小型套装饰品摆件仿真1.5米
双12礼遇季
此商品12.19开卖，请立即购买
价格 ¥128.00
礼遇价 ¥55.00 双旦礼遇季
淘金币可抵扣商品价格2%
运费 浙江金华 至 珠海 快递: 0.00
月销量 4.0万+ 累计评价 54549 送天猫积分 27
颜色分类
数量 1 件 库存1147件
立即购买
加入购物车
00:30 00:52
服务承诺 正品保证 极速退款 赠运费险 七天无理由退换
支付方式
商品详情 累计评价 54549 手机购买
品牌名称：华驰
产品参数：
品牌: 华驰
尺寸: 141cm(含)-170cm(含)
颜色分类: 1米含52个配饰 1....
毛重: 2.2KG
光源类型: 发光
货号: SDS-001
包装体积: 0.04
商品详情 累计评价 54549 手机购买
与描述相符
4.9
便宜(447) 质量好(393) 快递不错(157) 态度不错(103)
做工不错(40) 颜色漂亮(32) 质量一般(34)
全部 追评 (934) 图片 (9343)
有内容 按默认
女儿想要节日的气氛就给买了个，虽然安装的时候树枝有点扎手，但装好后确实漂亮，尤其晚上把彩灯打开后超级美，女女非常喜欢。
颜色分类：1.5米
h***1
解释：感谢您对我们的支持与信赖，您简短的几句话胜过无数的赞美，您的满意就是对我们最大的支持！希望给你带去的不仅仅是一份舒适分感觉，更是一份美丽的心情！

Why we need multimodality?

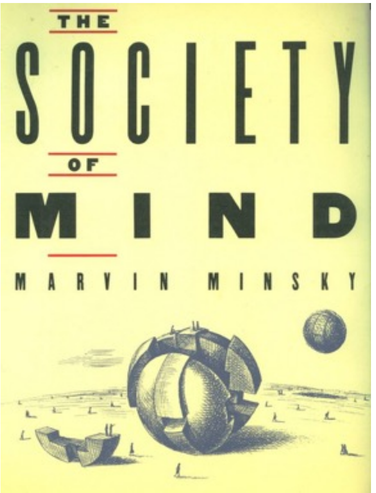
Cognitive and Knowledge Graph View



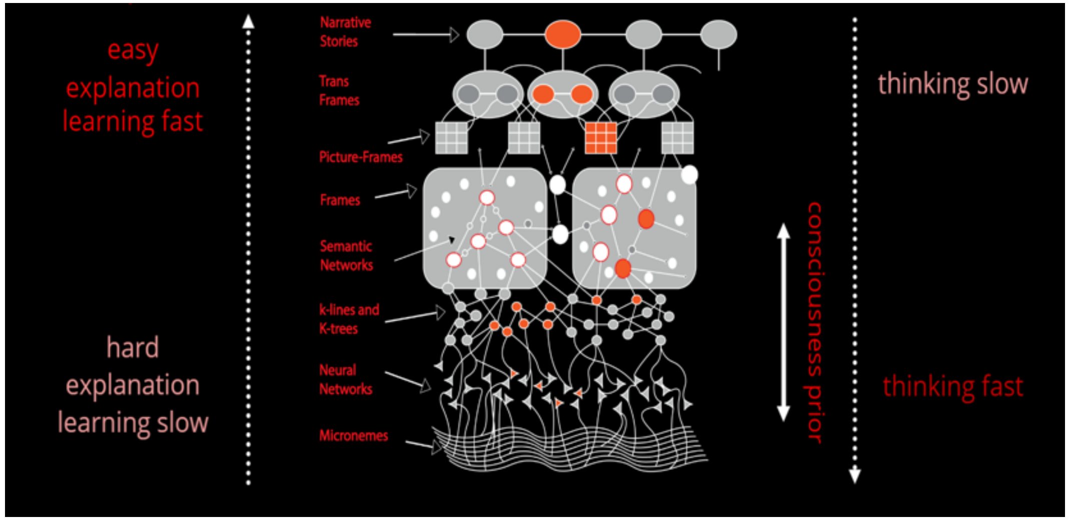
Yoshua Bengio
NeurIPS Keynote, 2019



From system 1 DL to system 2 DL



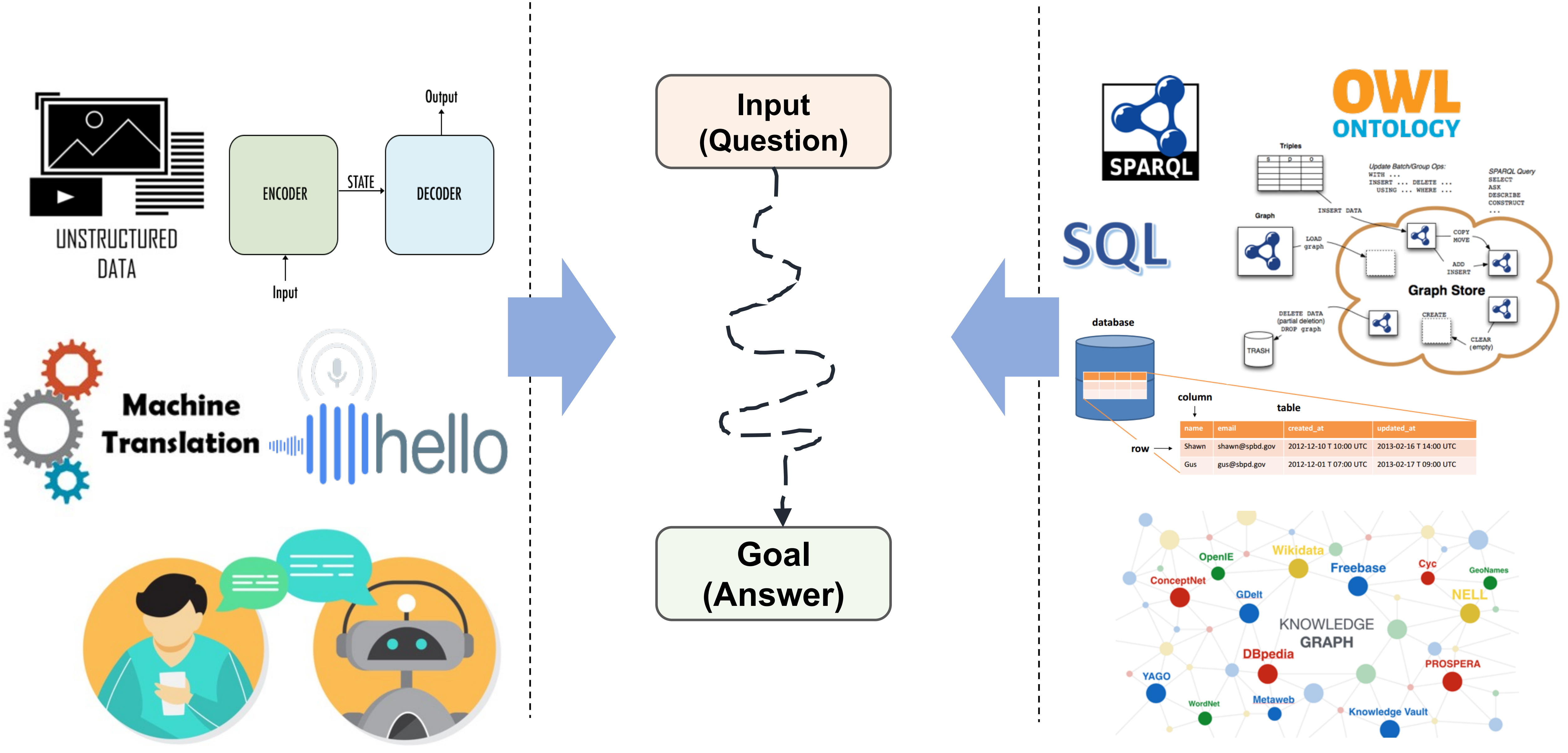
Marvin Minsky
The Society of Mind, 1986

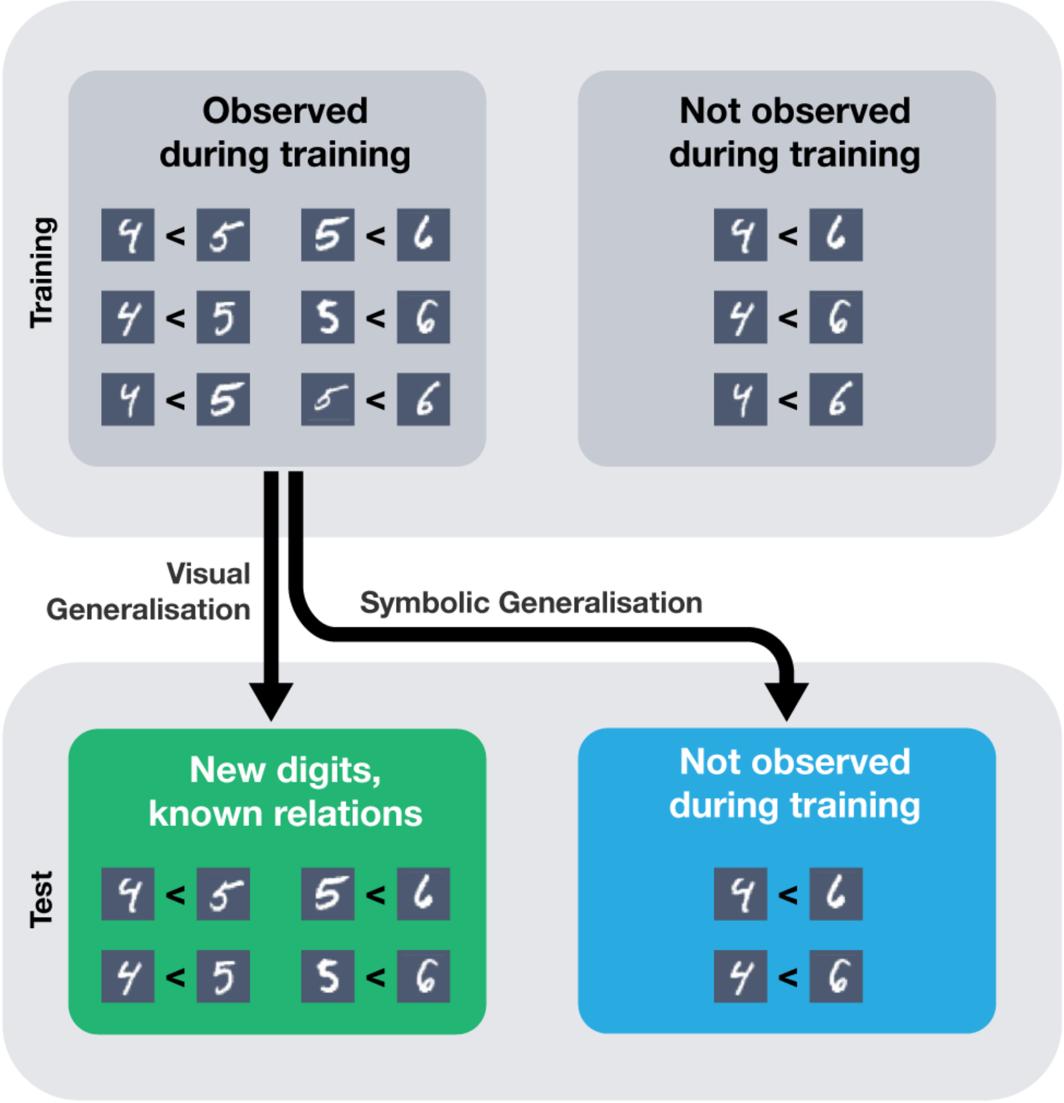


Framework for representing knowledge

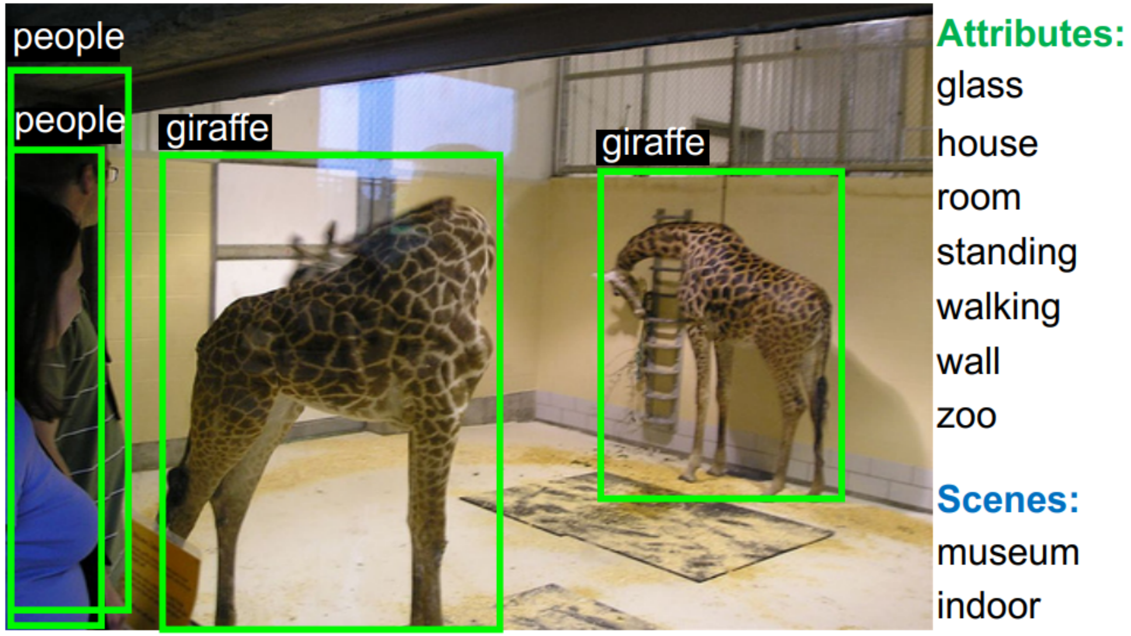
Neural (System1)

Symbolic (System2)





Visual generalisation vs. Symbolic generalisation



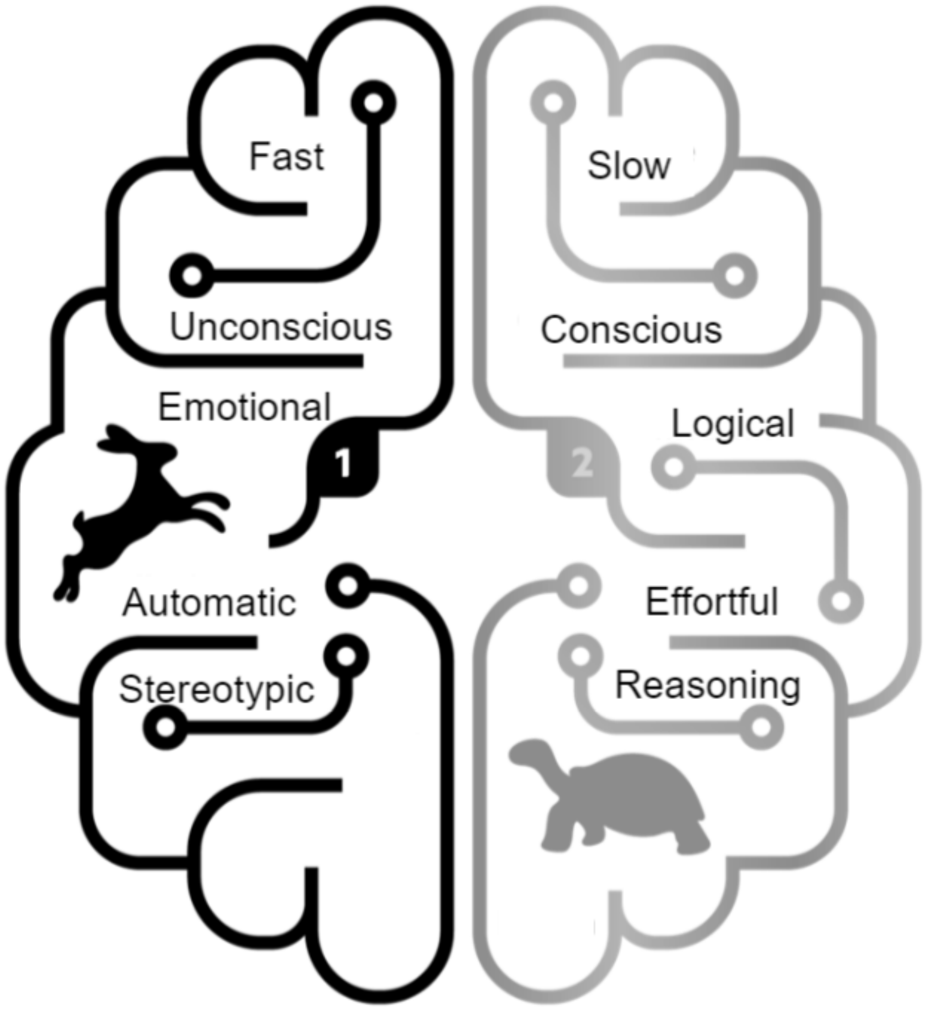
Visual Question: How many giraffes are there in the image?
Answer: Two.

Common-Sense Question: Is this image related to zoology?
Answer: Yes. Reason: Object/Giraffe --> Herbivorous animals --> Animal --> Zoology; Attribute/Zoo --> Zoology.

KB-Knowledge Question: What are the common properties between the animal in this image and zebra?
Answer: Herbivorous animals; Animals; Megafauna of Africa.

VQA, Commonsense QA, KBQA， and Machine Reading Comprehension

Cognitive Theory



Knowledge Graph Perspective

Neural (system1) are

- powerful for some problems
- robust to data noise
- hard to understand or explain
- poor at symbol manipulation
- unclear how to effectively use background knowledge

Symbolic (system2) are

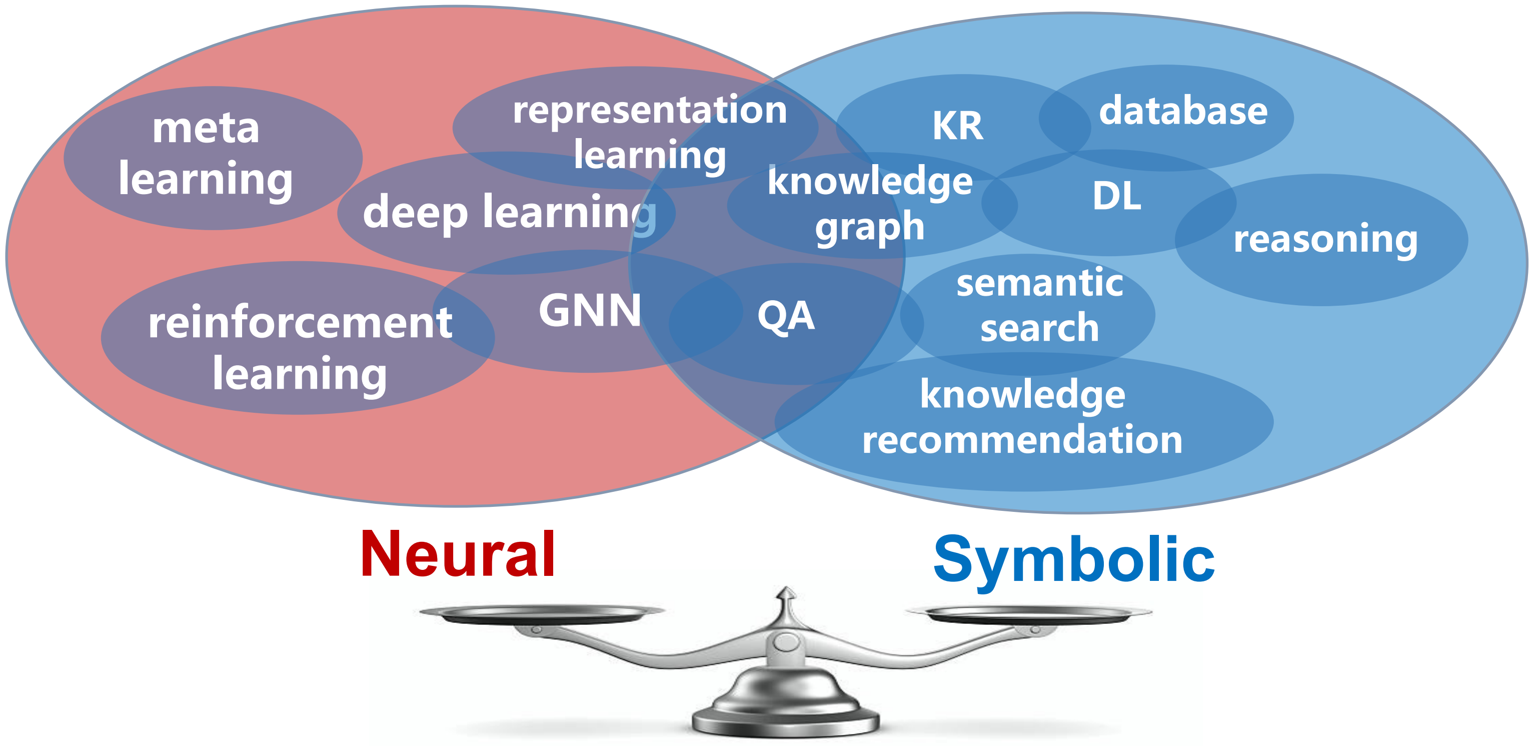
- Usually poor regarding machine learning problems
- Intolerant to data noise
- Easy to understand and assess by a human
- Good at symbol manipulation
- Designed to work with background knowledge

“神经+符号” ：从知识图谱角度看认知推理的发展,《中国计算机学会通讯》，2020年第16卷第8期

Neural+Symbolic:

- powerful machine learning paradigm
- robust to data noise
- easy to understand and assess by humans
- good at symbol manipulation
- work seamlessly with background knowledge

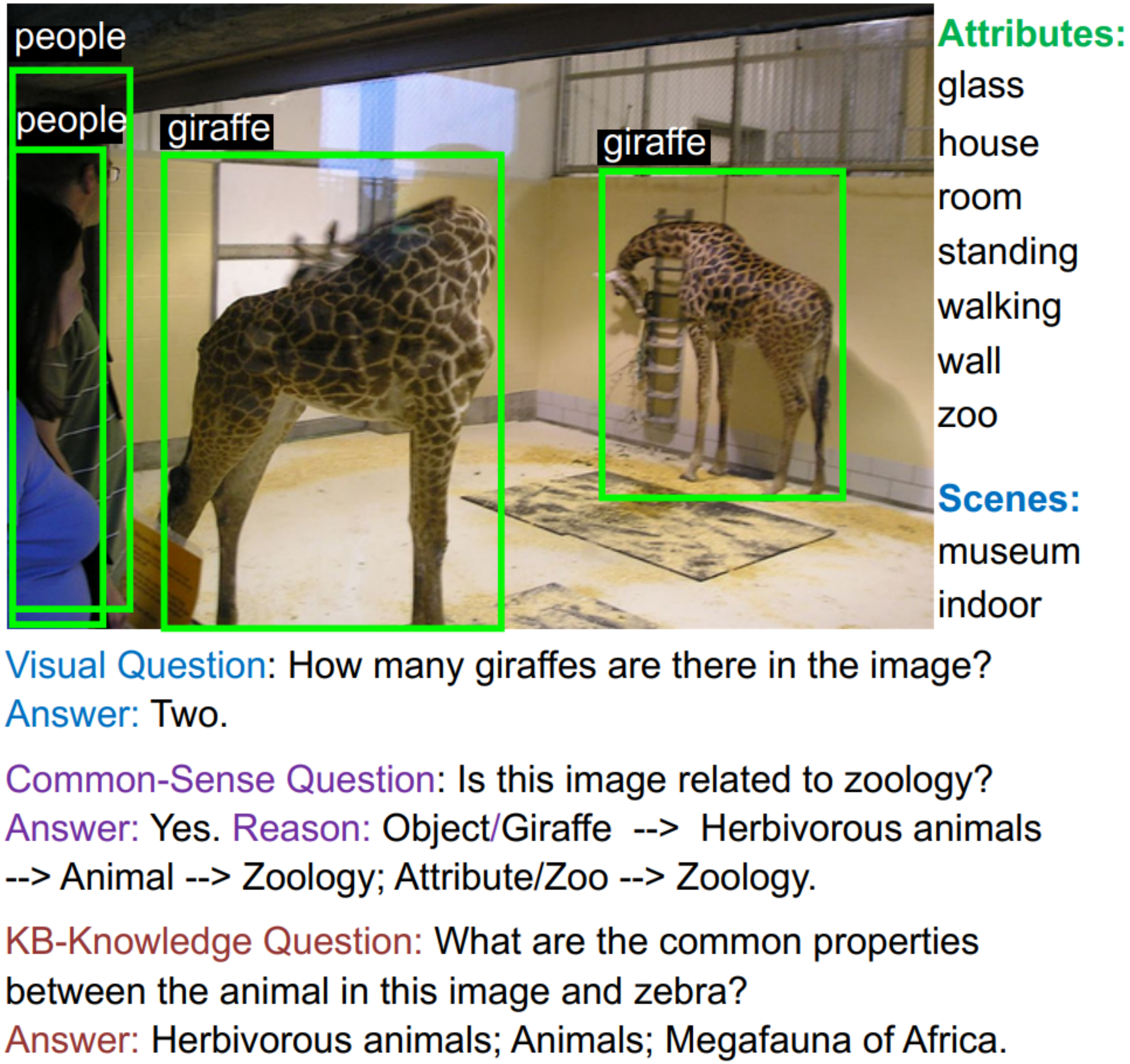
HOW TO Multimodal Knowledge Graph ?



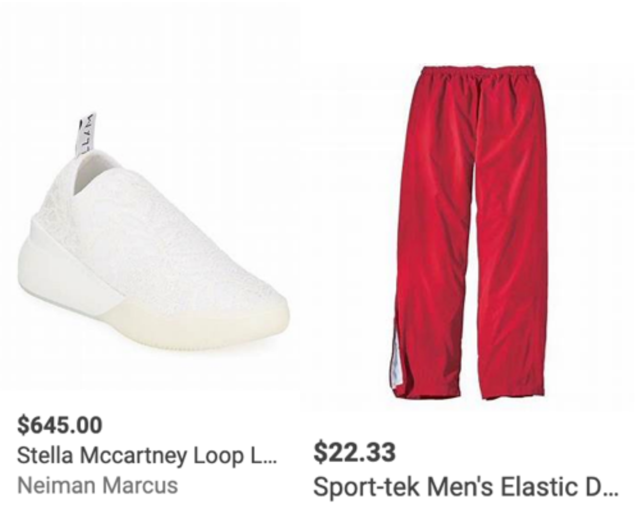
Application View

More cross-modal relations, more details and more answers

Cross-modal entity grounding

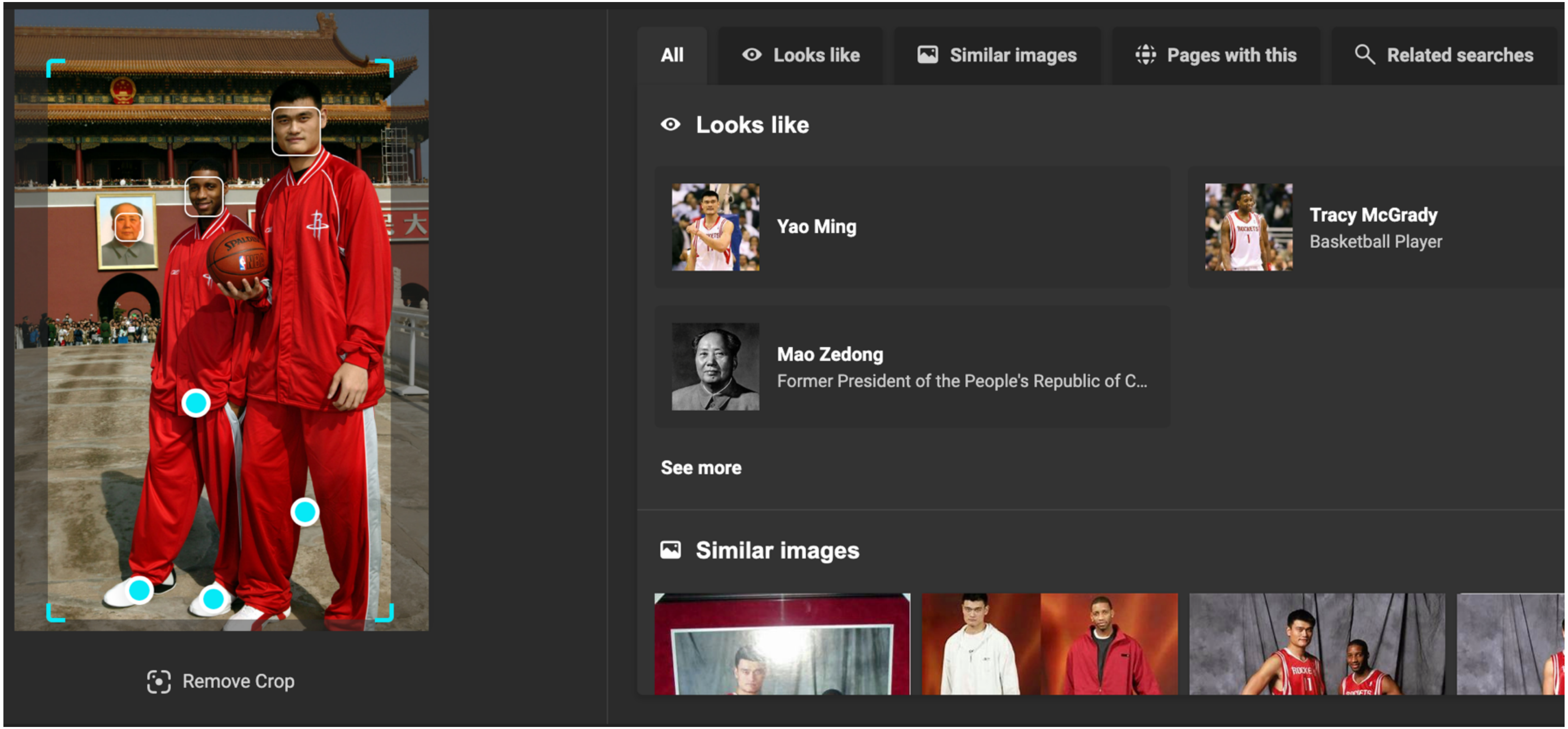


VQA, Complicated scene understanding

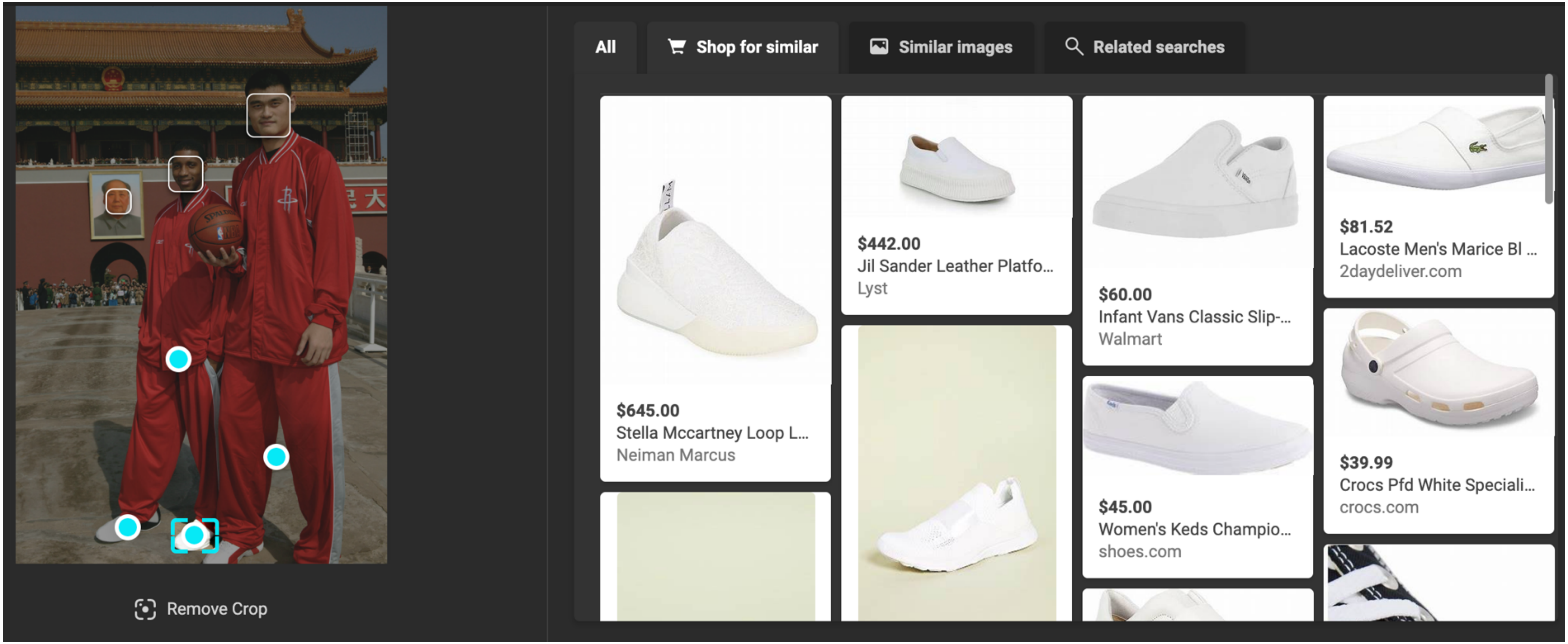


HOUSTON ROCKETS

Yao Ming and Tracy McGrady of the Houston Rockets visit Beijing in 2004 bing


All
Looks like
Similar images
Pages with this
Related searches
Looks like
Yao Ming
Tracy McGrady
Basketball Player
Mao Zedong
Former President of the People's Republic of C...
See more
Similar images
Remove Crop

bing


All
Shop for similar
Similar images
Related searches
$645.00
Stella Mccartney Loop L...
Neiman Marcus
$442.00
Jil Sander Leather Platfo...
Lyst
$60.00
Infant Vans Classic Slip-...
Walmart
$45.00
Women's Keds Champio...
shoes.com
$81.52
Lacoste Men's Marice Bl ...
2daydeliver.com
$39.99
Crocs Pfd White Speciali...
crocs.com
Remove Crop

Cross-modal Disambiguation:
Heterogeneous in modal, but correlated in semantic

Result：
姚明（100）
威尔史密斯（39）
梅兰芳（36）

Visual Entity Disambiguation

刘欢

这是一个多义词，请在下列义项上选择浏览（共14个义项）

- 刘欢：中国内地流行音乐家
- 刘欢：广东省广州市中级人民法院助理审判员
- 刘欢：中国足球运动员
- 刘欢：长虹街道办事处副主任
- 刘欢：湖南发展研究中心研究员联络处副主任
- 刘欢：清华大学环境学院副教授
- 刘欢：清华大学教师
- 刘欢：象棋棋手
- 刘欢：矿大（北京）管院第十二届研究生会副主席
- 刘欢：苏州东吴队球员
- 刘欢：中国大陆男演员
- 刘欢：扣篮王刘欢
- 刘欢：全国技术能手

刘欢在美国超市被偶遇，买8美元面包，
爽快接过纸笔给网友签名

Textual Entity Disambiguation

基本信息

中文名	刘欢	毕业院校	国际关系学院法国文学专业
外文名	Liu Huan	经纪公司	百娱传媒股份有限公司
别　名	欢哥	代表作品	少年壮志不言愁、弯弯的月亮、心中的太阳、千万次
国　籍	中国		的问、这一拜、好汉歌、从头再来、凤凰于飞
民　族	汉族	主要成就	CCTV MTV音乐盛典最受欢迎男歌手
星　座	处女座		《音乐风云榜》终身成就奖
血　型	O型		北艺协会电视剧优秀音乐创作奖
身　高	173cm		第十届华语歌曲"榜中榜"之"评委会特别奖"
出生地	天津		第四届中国金唱片"最佳流行专辑"
出生日期	1963年8月26日	生　肖	兔
职　业	歌唱家、音乐人、词曲创作人、大学音乐教授		

Huan Liu

Computer scientist

Huan Liu is a computer scientist at Arizona State University in Tempe, Arizona. He was named a Fellow of the Institute of Electrical and Electronics Engineers in 2012 for his contributions to feature selection in data mining and knowledge discovery. Wikipedia

What is Multimodal Knowledge Graph?

MMKG

ESP Game
2004

- Labeling Images with a Computer Game

iM
2008，2009

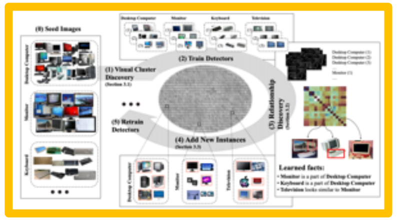
- Interlinking Multimedia
- Apply Linked Data Principles to Multimedia Fragments

ImageNet,
Visipedia
2009,2010

built upon the backbone of the WordNet

IMAGENET

VISIpedia
Encyclopedia for Global Change



- Automatically extracting
- Semi-supervised learning
- Discovers common sense relationships

NEIL: Image
Knowledge
Miner
2013

IMGpedia
2015,2017

Semantic Web formats

IMGpedia

Richpedia

Still many unsolved problems

Multimodal
Knowledge
Graph
2019

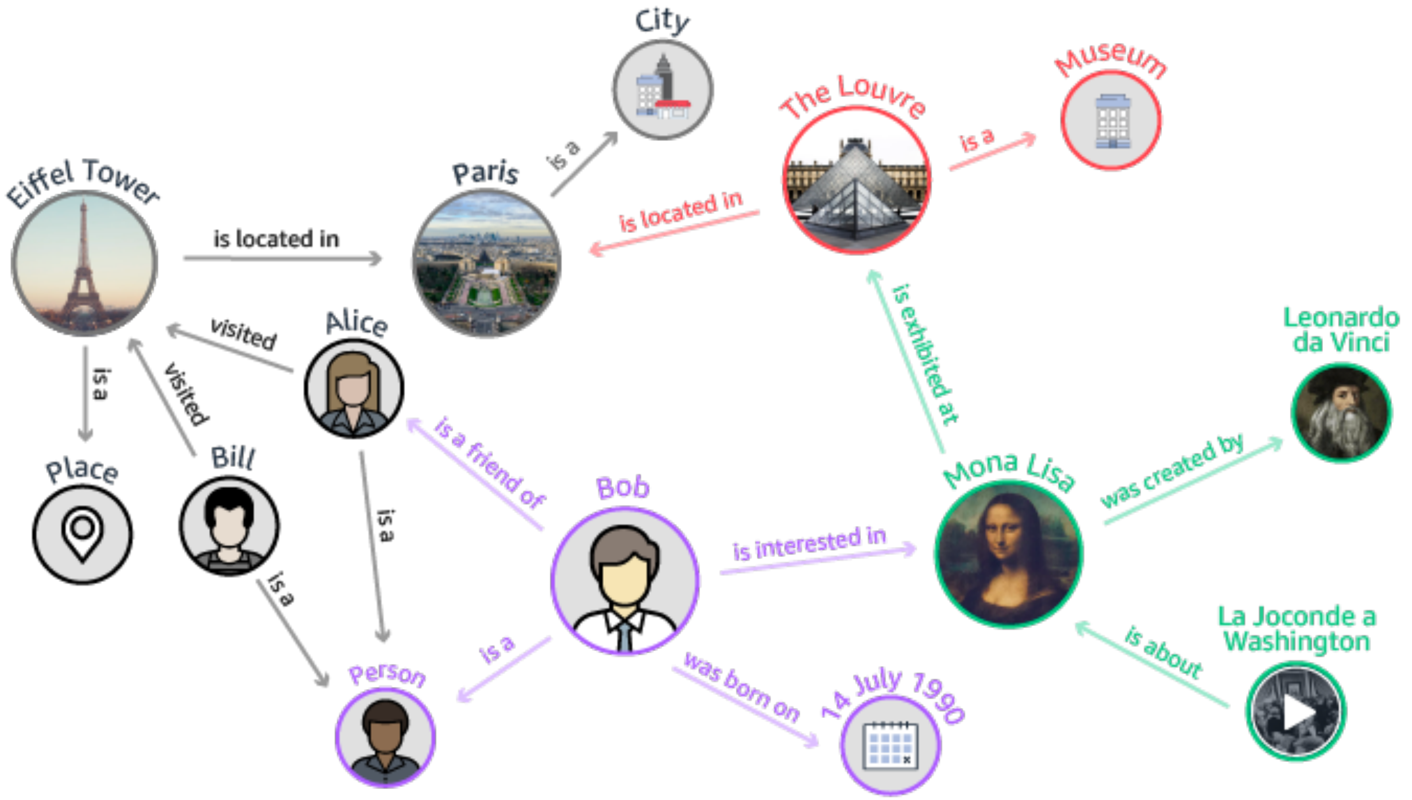
Emerging Multimodal KG Work

Node:

- Image entity
- Text entity
- Visual concept
- Textual concept

Relation:

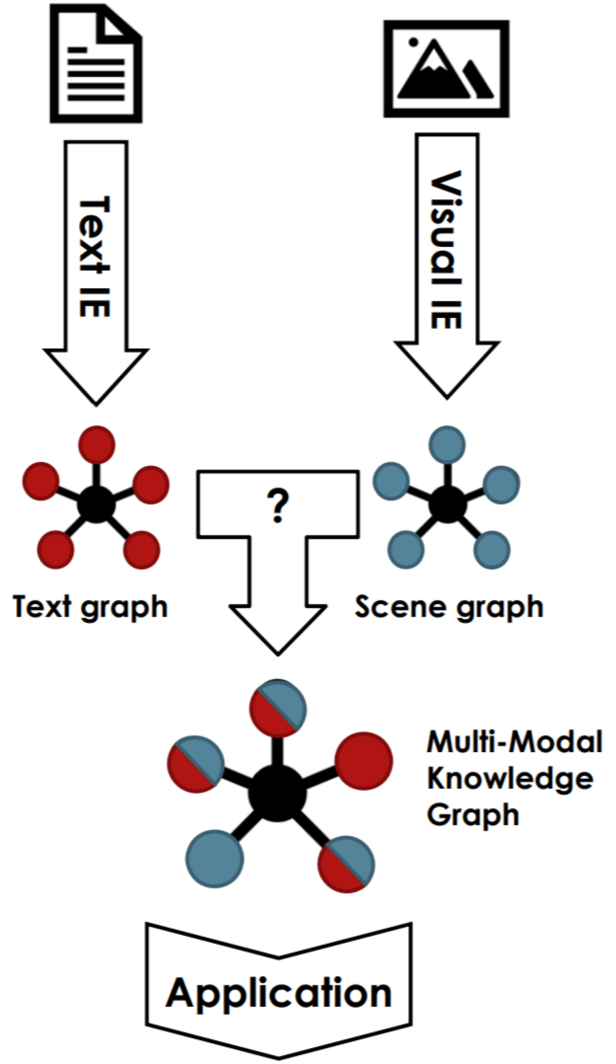
- is-a
- has-visual-object
- meta-of
- has-tag
- co-locate-with



Dihong Gong , Daisy Zhe Wang
Towards Building Large-Scale Multimodal Knowledge Bases

Emerging Multimodal KG Work

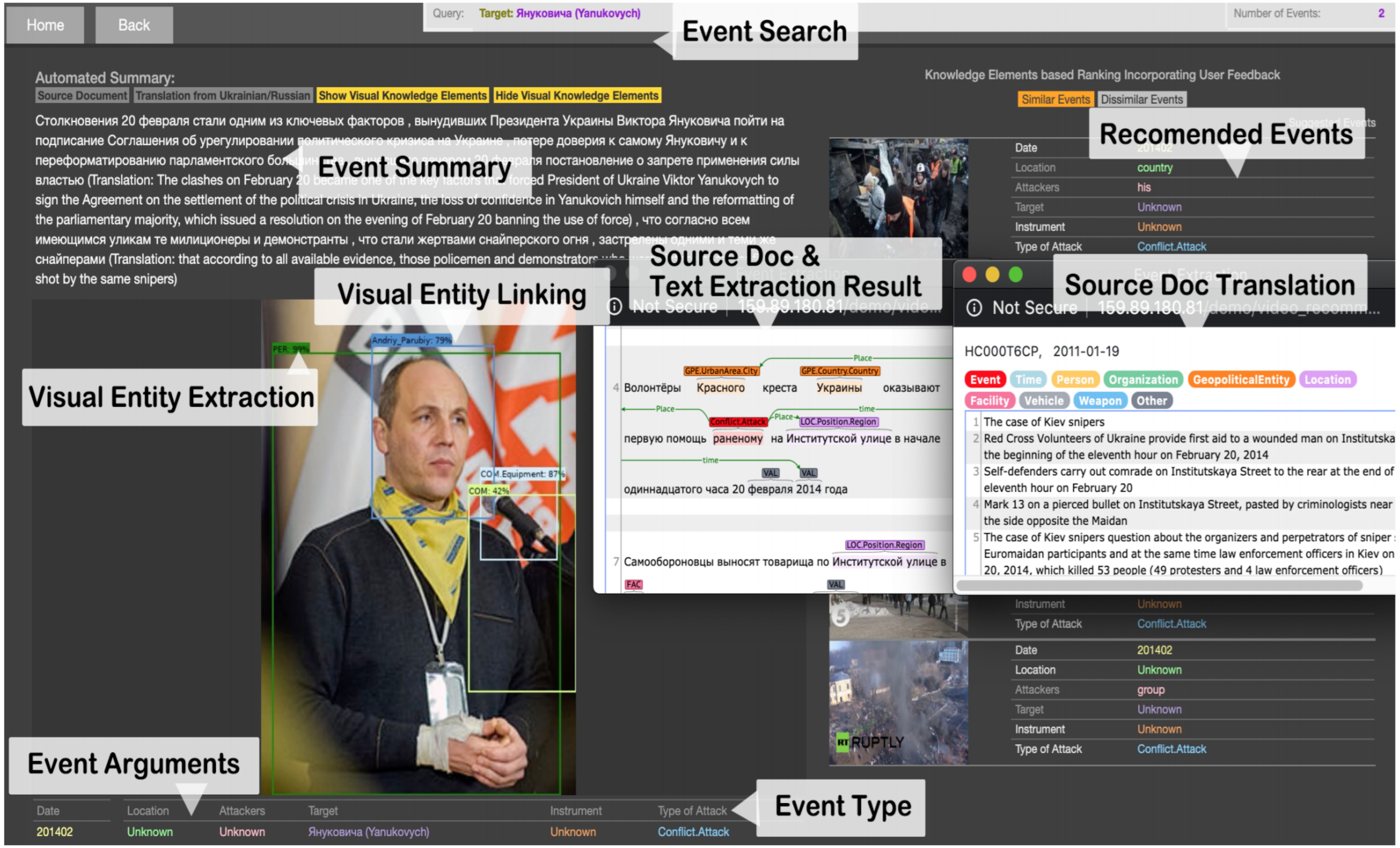
- Challenges:
  - Parsing text to structured semantic graph
  - Parsing images/videos to structures
  - Grounding event/entities across modalities
  - Multimodal argument role
- Applications
  - Story Generation and Summarization
  - Question Answering
  - Commonsense Discovery



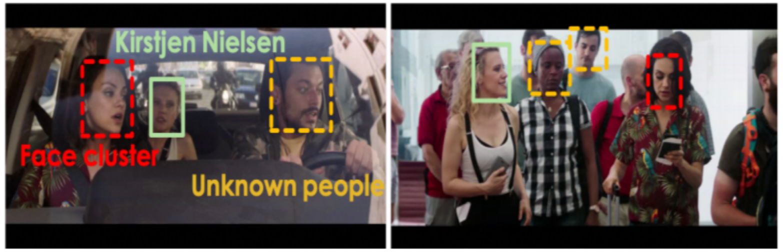
Shih-Fu Chang, Alireza Zareian, Hassan Akbari, Brian Chen, Heng Ji, Spencer Whitehead, Manling Li

Multimodal Knowledge Graphs: Automatic Extraction & Applications

Emerging Multimodal KG Work

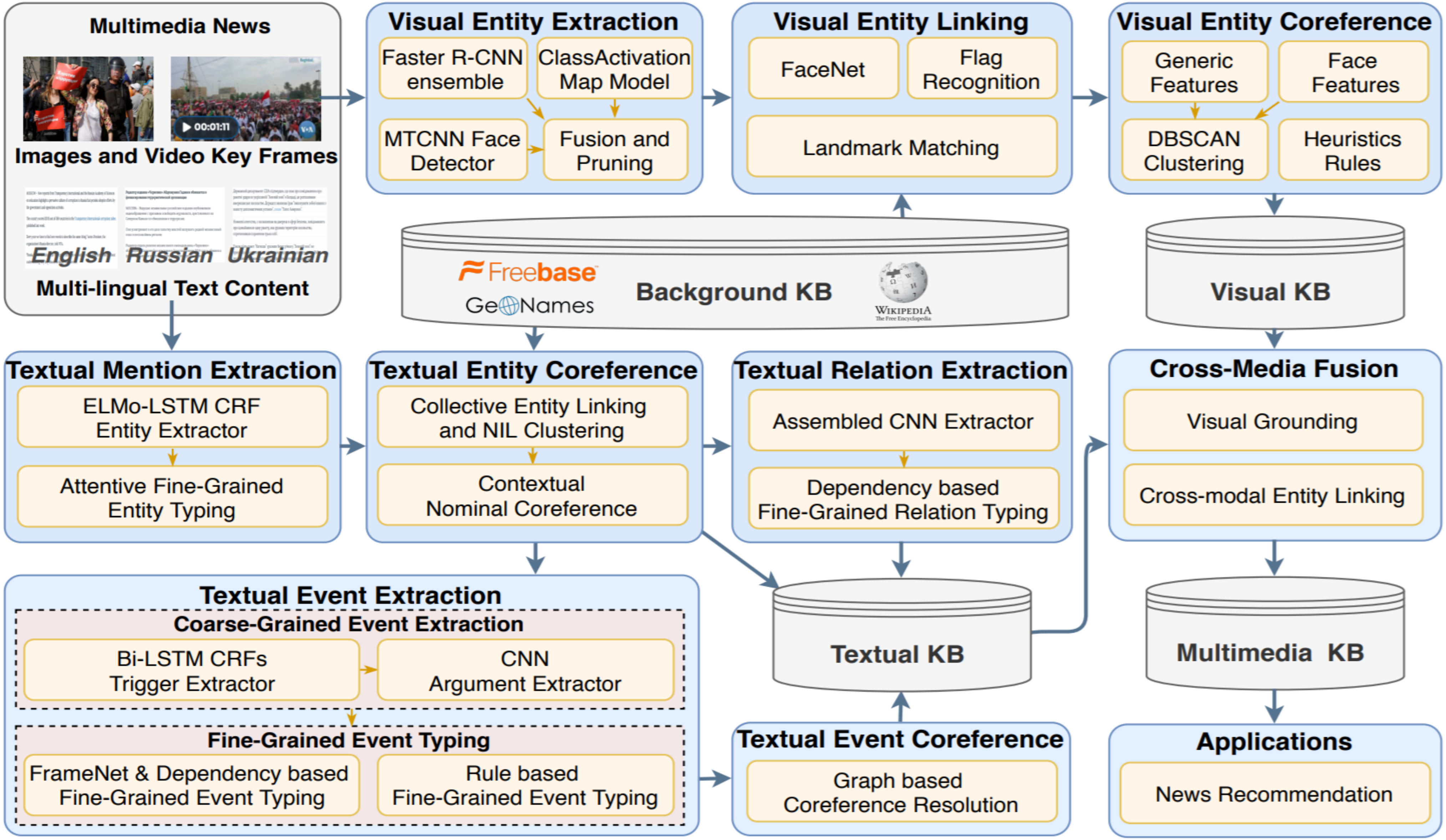


	Coarse-grained Types	Fine-grained Types
Entity	7	187
Relation	23	61
Event	47	144



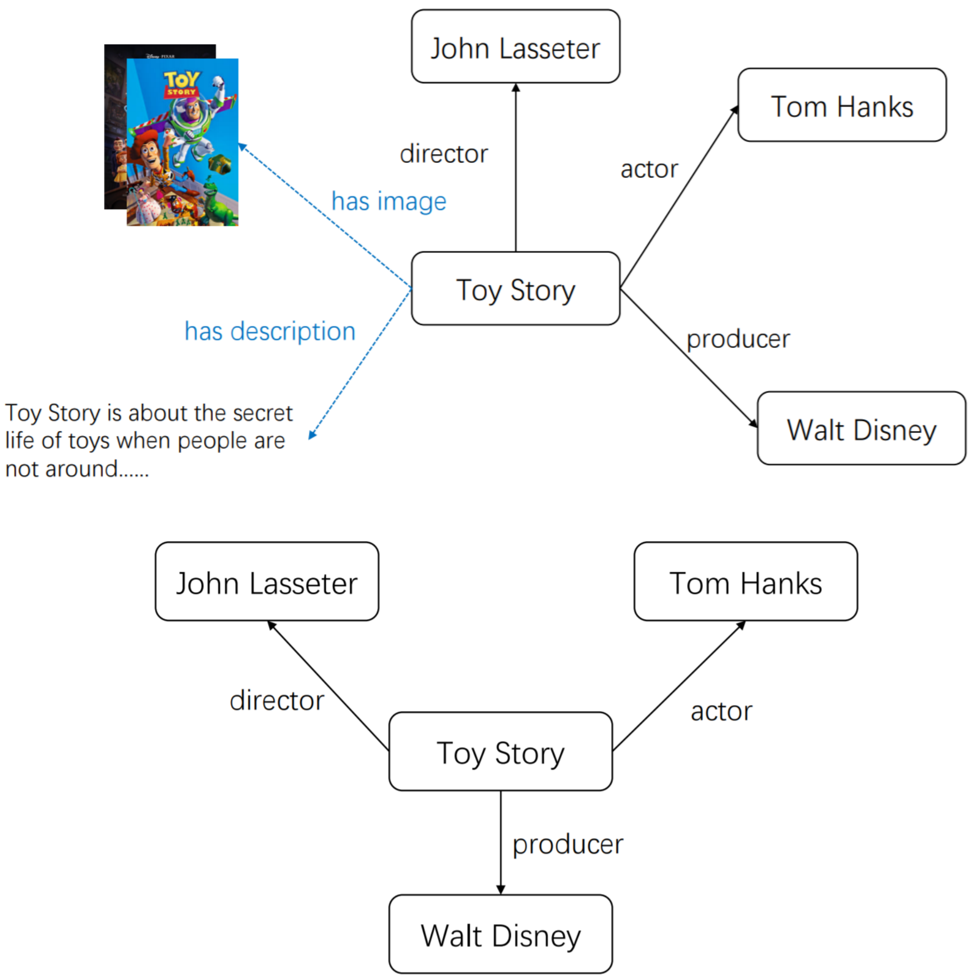
Li, Manling, et al. "Gaia: A fine-grained multimedia knowledge extraction system." *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. (ACL 2020).

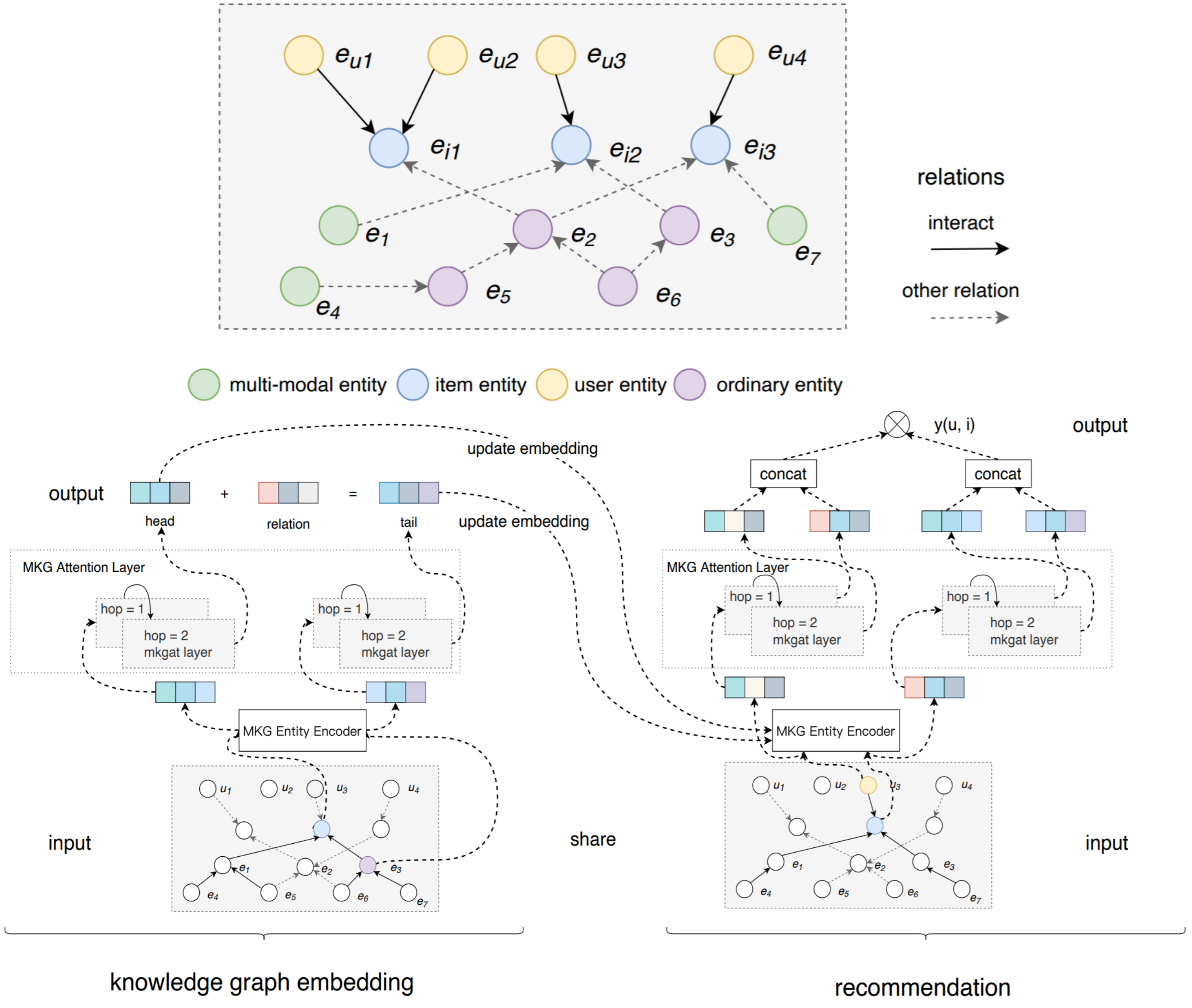
Emerging Multimodal KG Work



Li, Manling, et al. "Gaia: A fine-grained multimedia knowledge extraction system." *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. (ACL 2020).

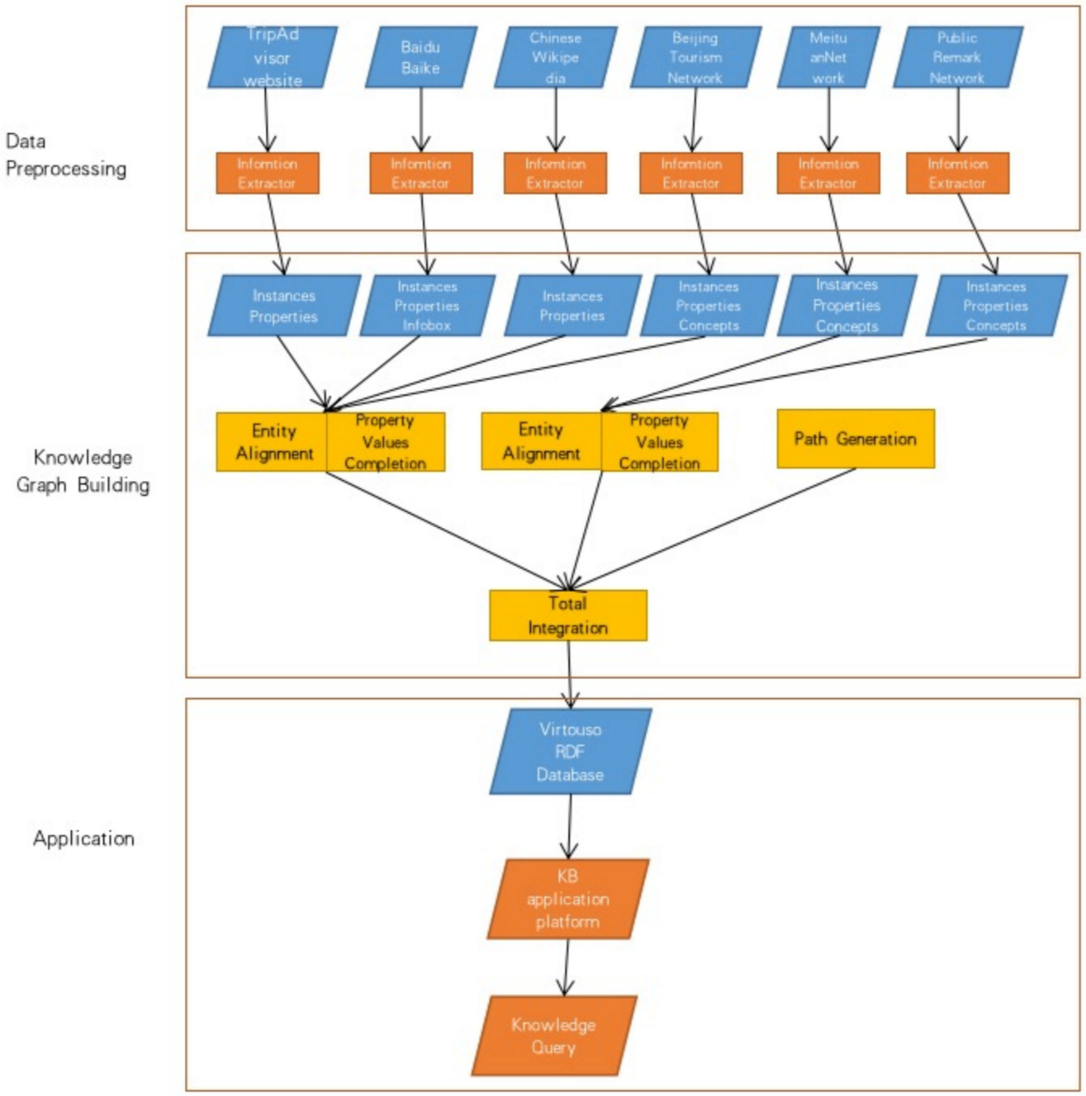
Emerging Multimodal KG Work

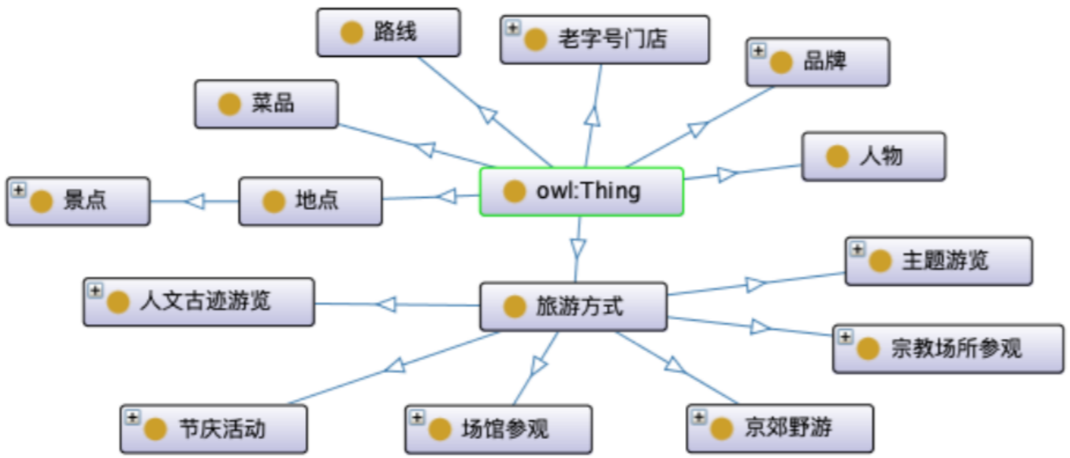




Sun, Rui, et al. "Multi-modal Knowledge Graphs for Recommender Systems." *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM 2020).*

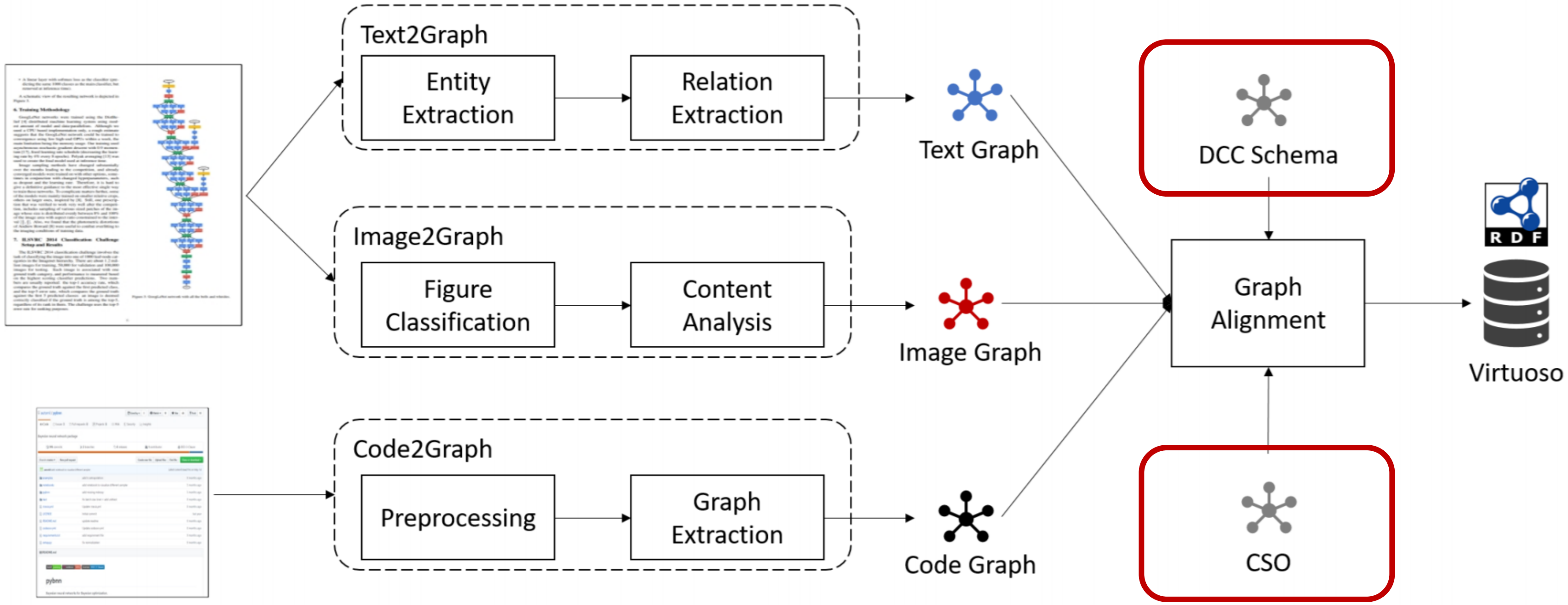
Emerging Multimodal KG Work

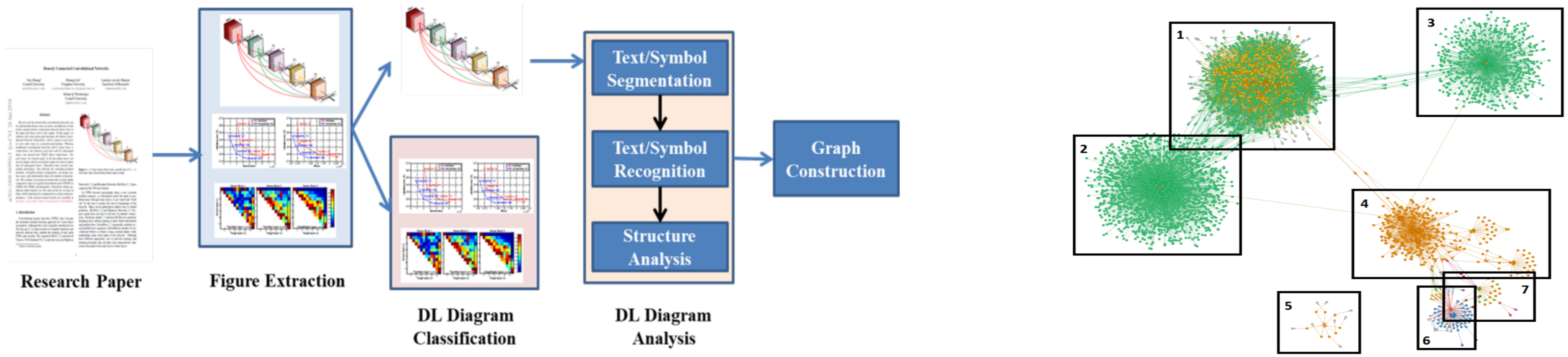




Xie, Qinghua Wen, et al. “Construction of Multi-modal Chinese Tourism Knowledge Graph”

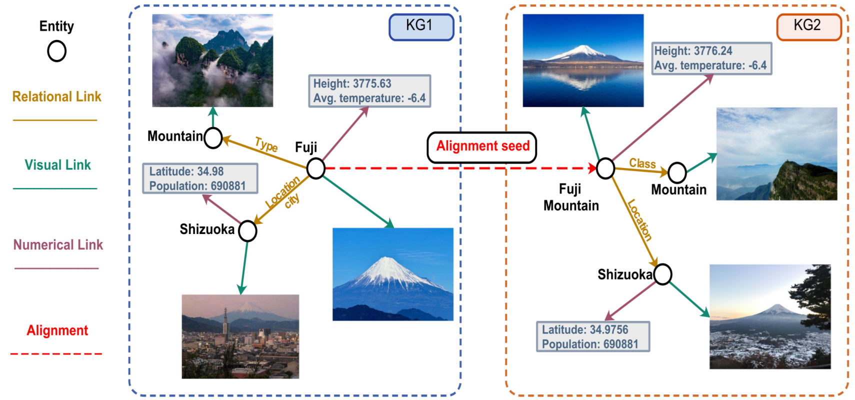
Emerging Multimodal KG Work





Kannan, Amar Viswanathan, et al. "Multimodal Knowledge Graph for Deep Learning Papers and Code." *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM 2020)*.

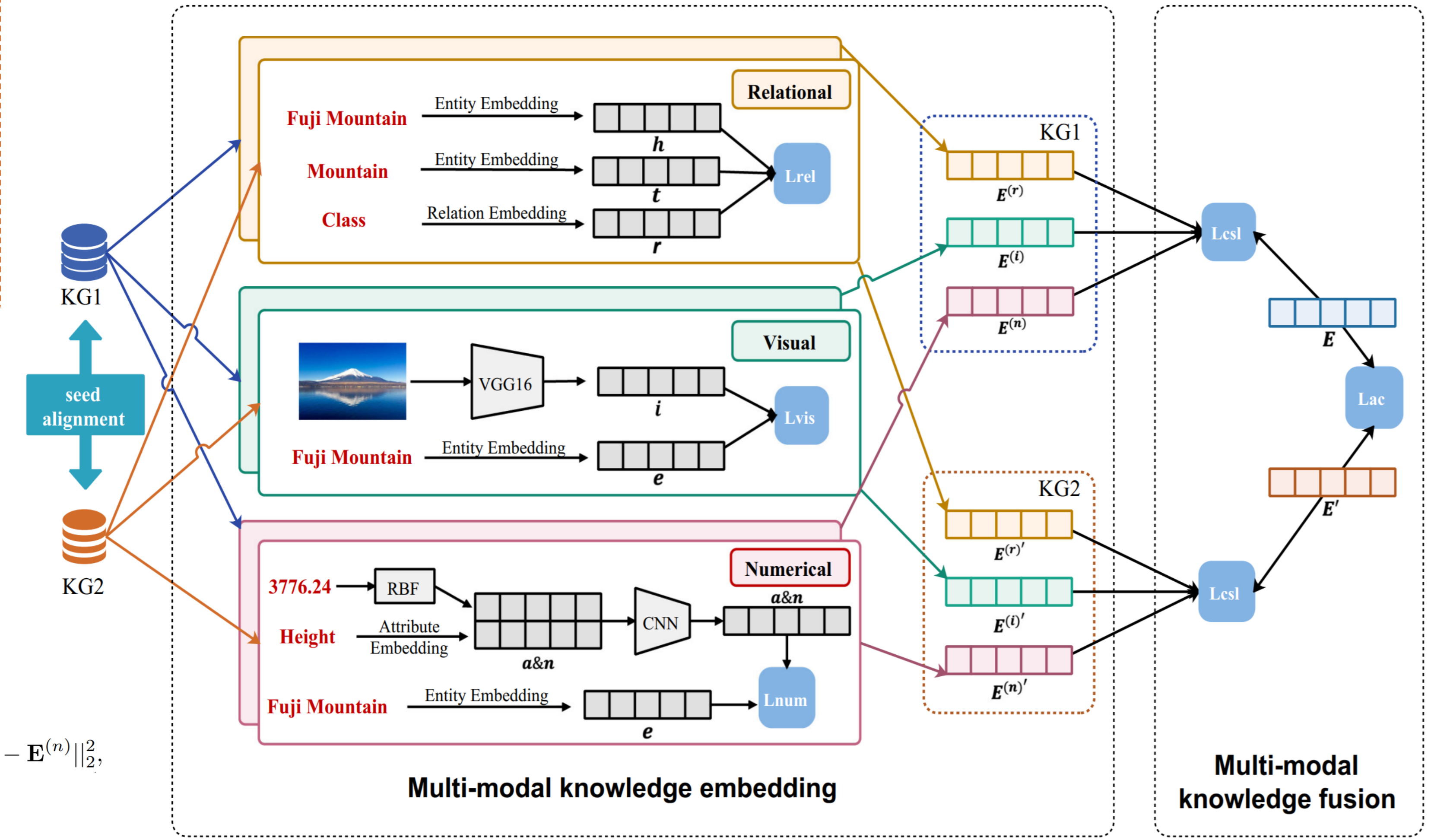
Emerging Multimodal KG Work



Datasets:
MMKG FB15K-DB15K and FB15K-YAGO15K

$$\mathcal{L} = -\sum_{\mathtt{t}\in\mathbf{T}} \log p(\mathtt{t} \mid \theta_1, \ldots, \theta_n).$$

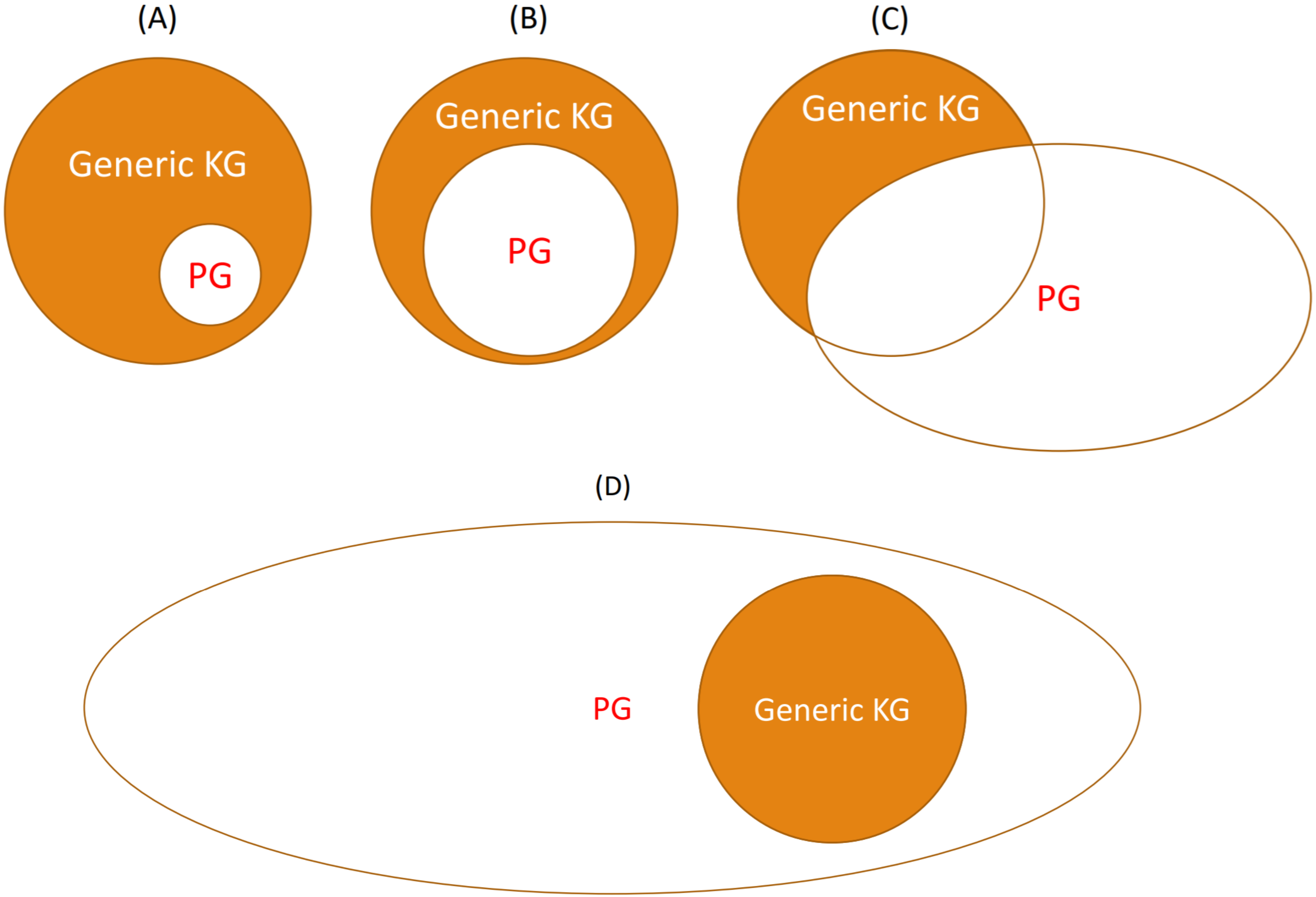
$$L_{csl}(\mathbf{E}, \mathbf{E}^{(r)}, \mathbf{E}^{(i)}, \mathbf{E}^{(n)}) = \alpha_1 ||\mathbf{E} - \mathbf{E}^{(r)}||_2^2 + \alpha_2 ||\mathbf{E} - \mathbf{E}^{(i)}||_2^2 + \alpha_3 ||\mathbf{E} - \mathbf{E}^{(n)}||_2^2,$$



Liu, Ye, et al. "MMKG: multi-modal knowledge graphs." *European Semantic Web Conference (ESWC 2019).*

Chen, Liyi, et al. "MMEA: Entity Alignment for Multi-modal Knowledge Graph." *International Conference on Knowledge Science, Engineering and Management (KSEM 2020).* (**Best Paper**)

Emerging Multimodal KG Work



Xin Luna Dong, Building a Broad Knowledge Graph for Products, 2019

Emerging Multimodal KG Work

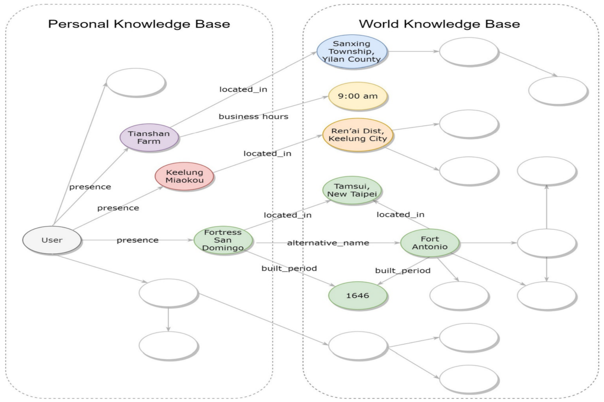
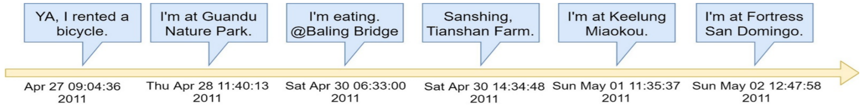
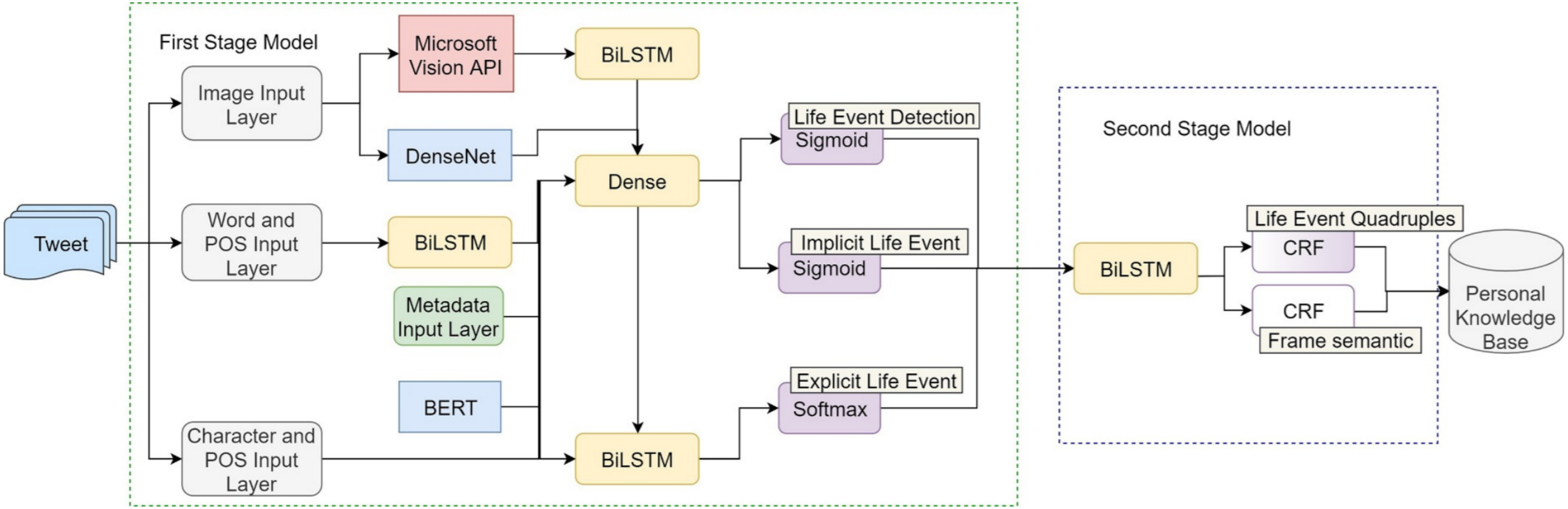


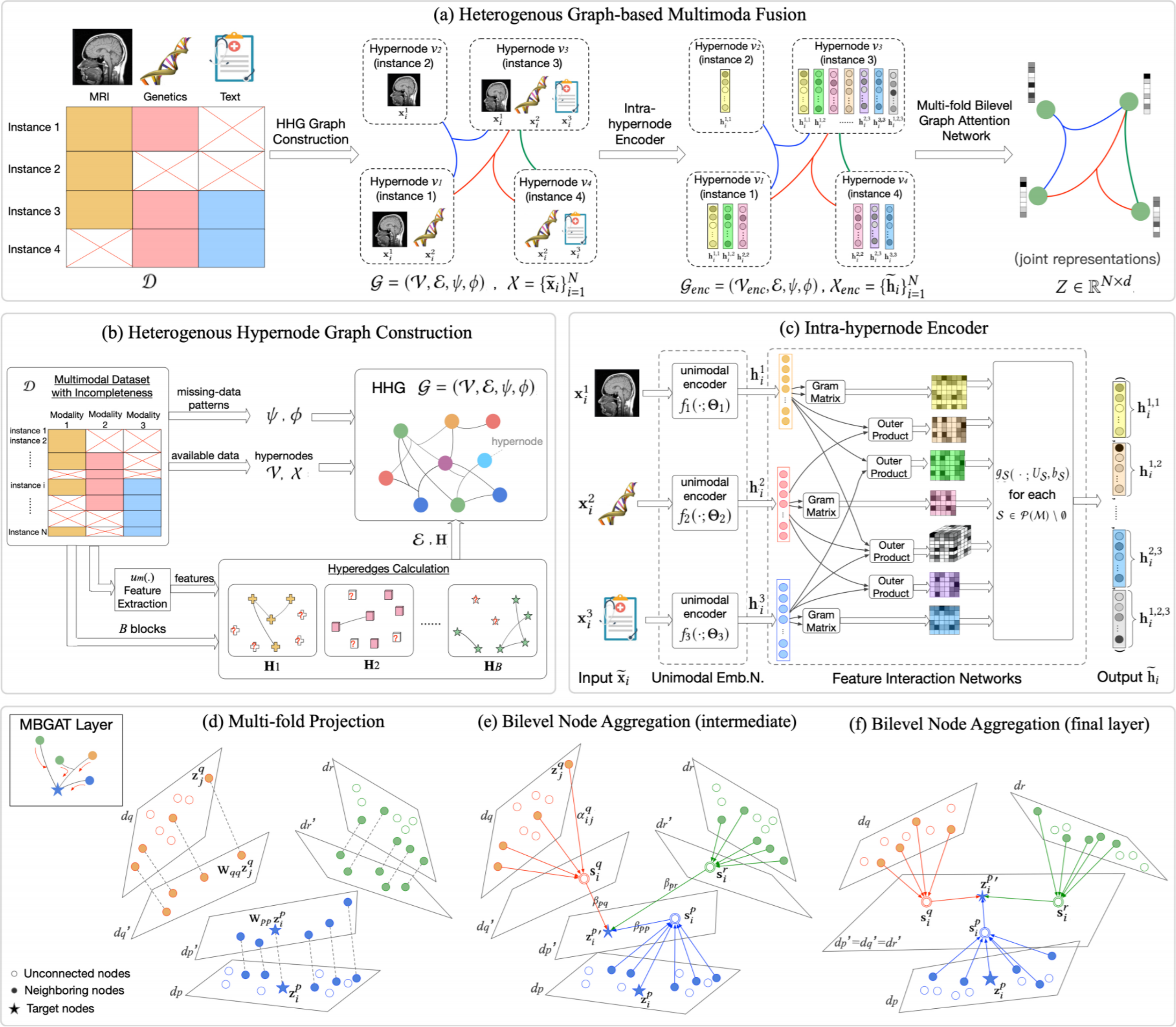
Fig. 1. An Example of social media post.

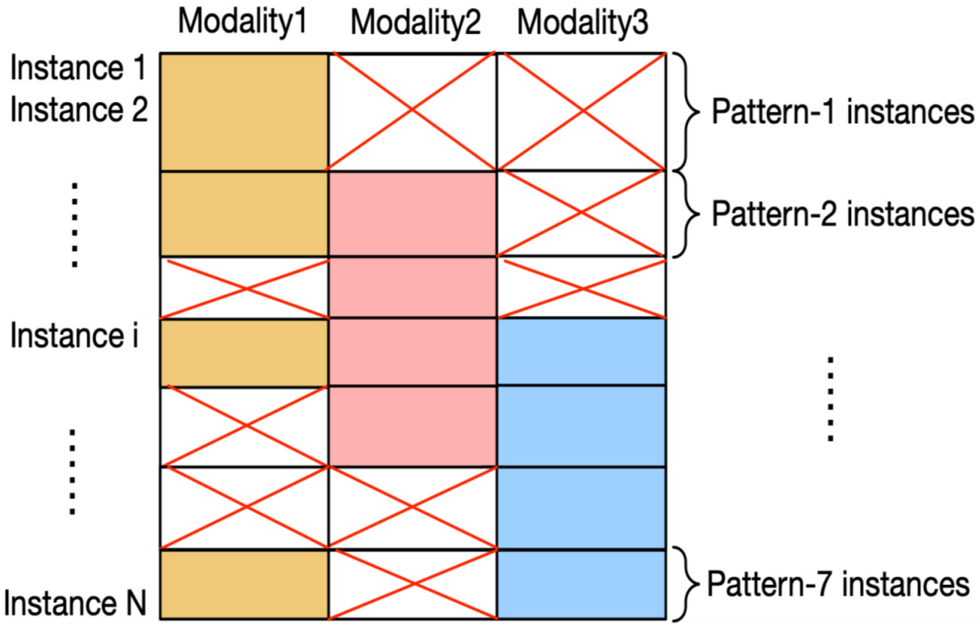




Yen, An-Zi, Hen-Hsen Huang, and Hsin-Hsi Chen. "Multimodal joint learning for personal knowledge base construction from Twitter-based lifelogs." *Information Processing & Management* (2019): 102148.

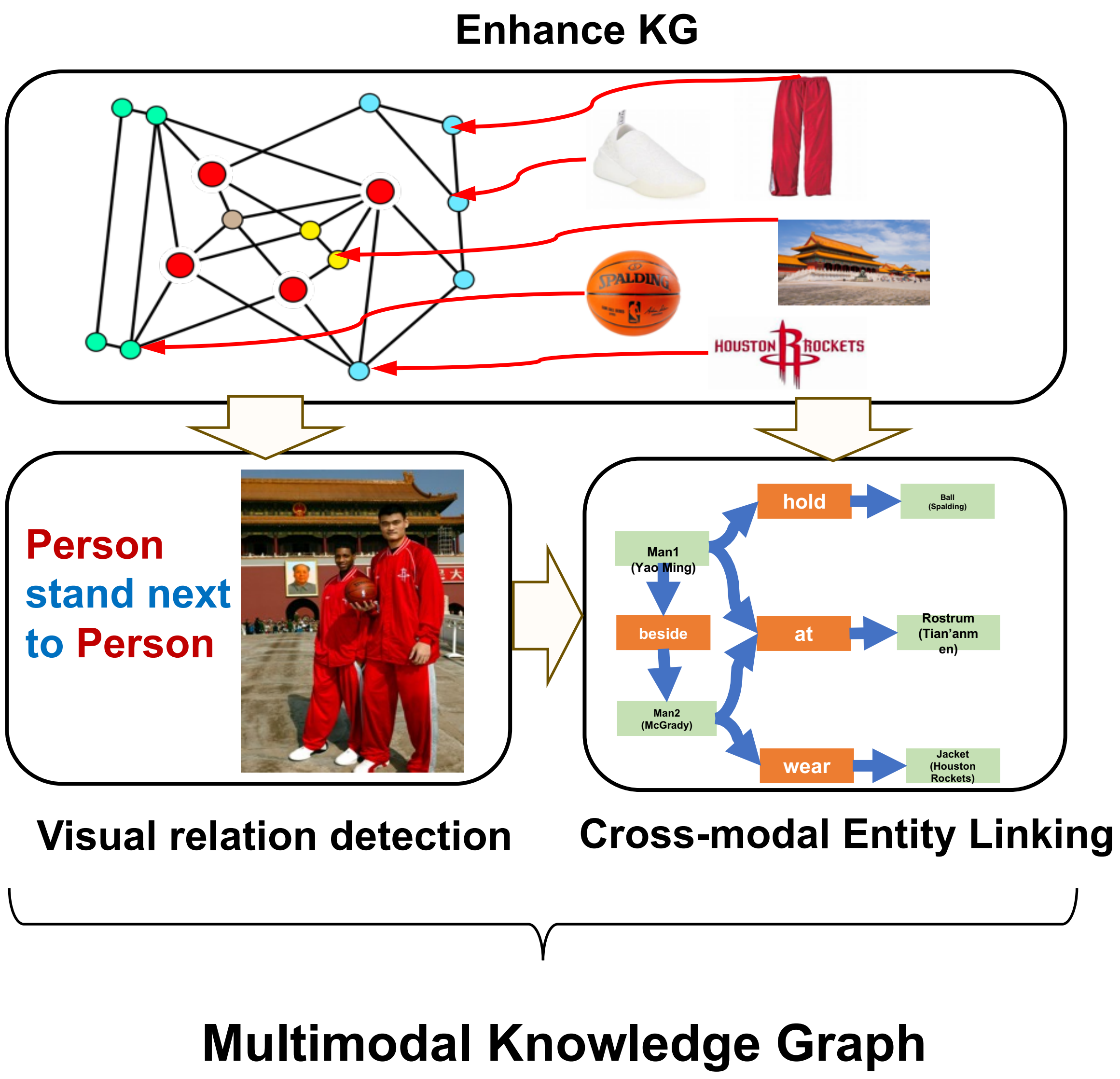
Emerging Multimodal KG Work

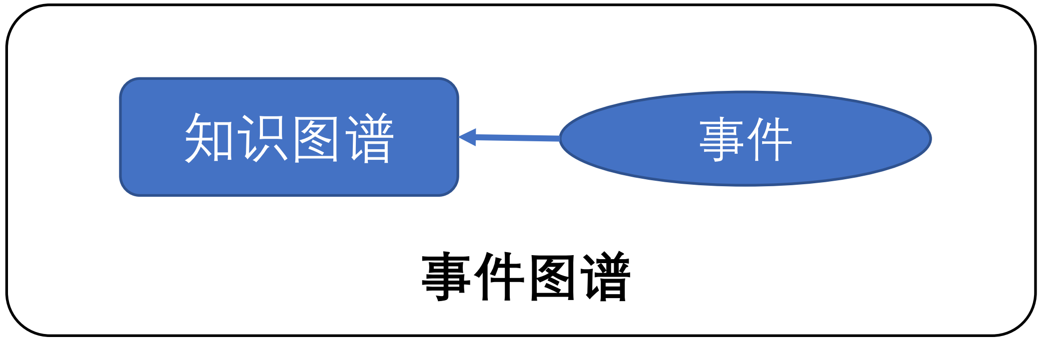


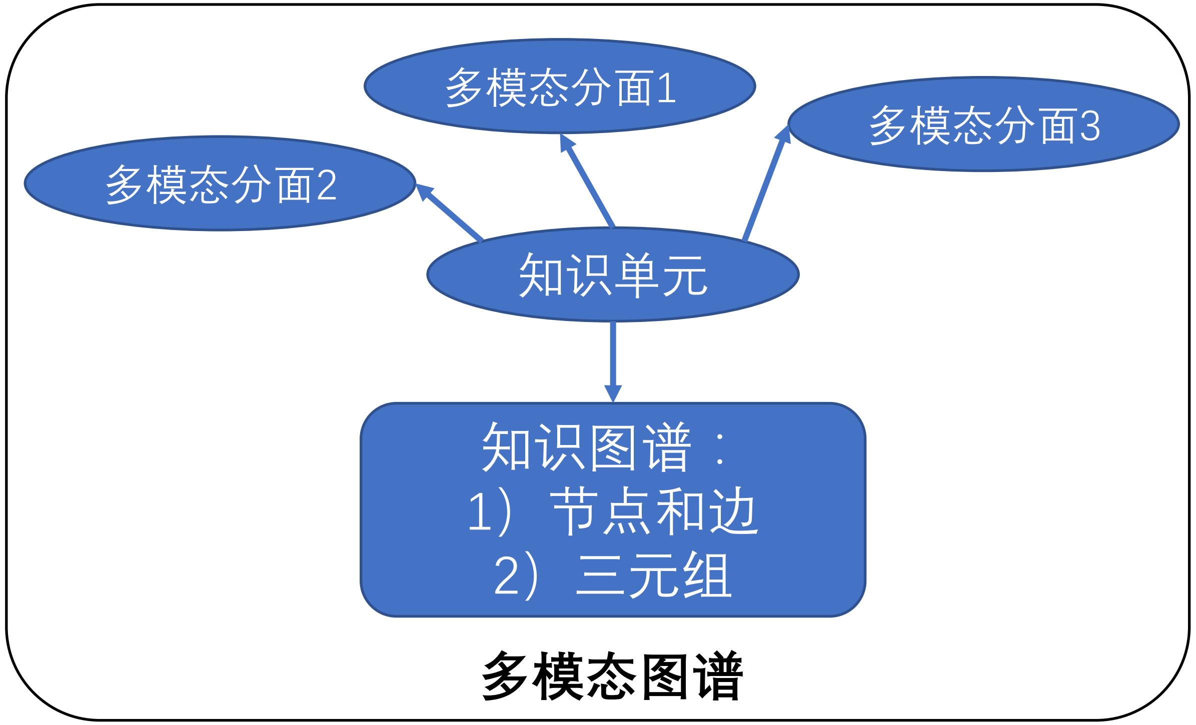


Jiayi Chen, Aidong Zhang “HGMF: Heterogeneous Graph-based Fusion for Multimodal Data with Incompleteness”, KDD 2020

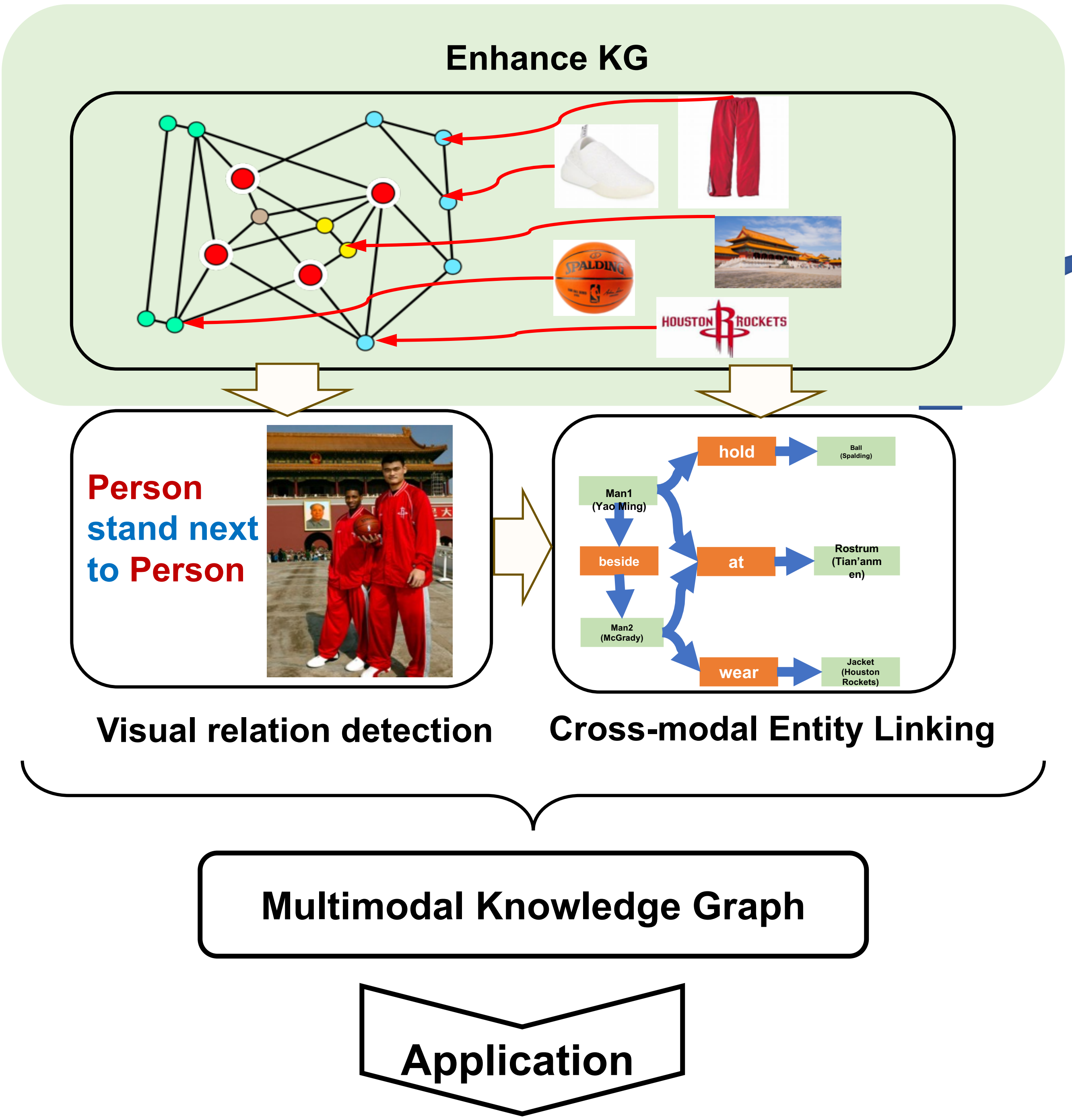
Our Multimodal Knowledge Graph





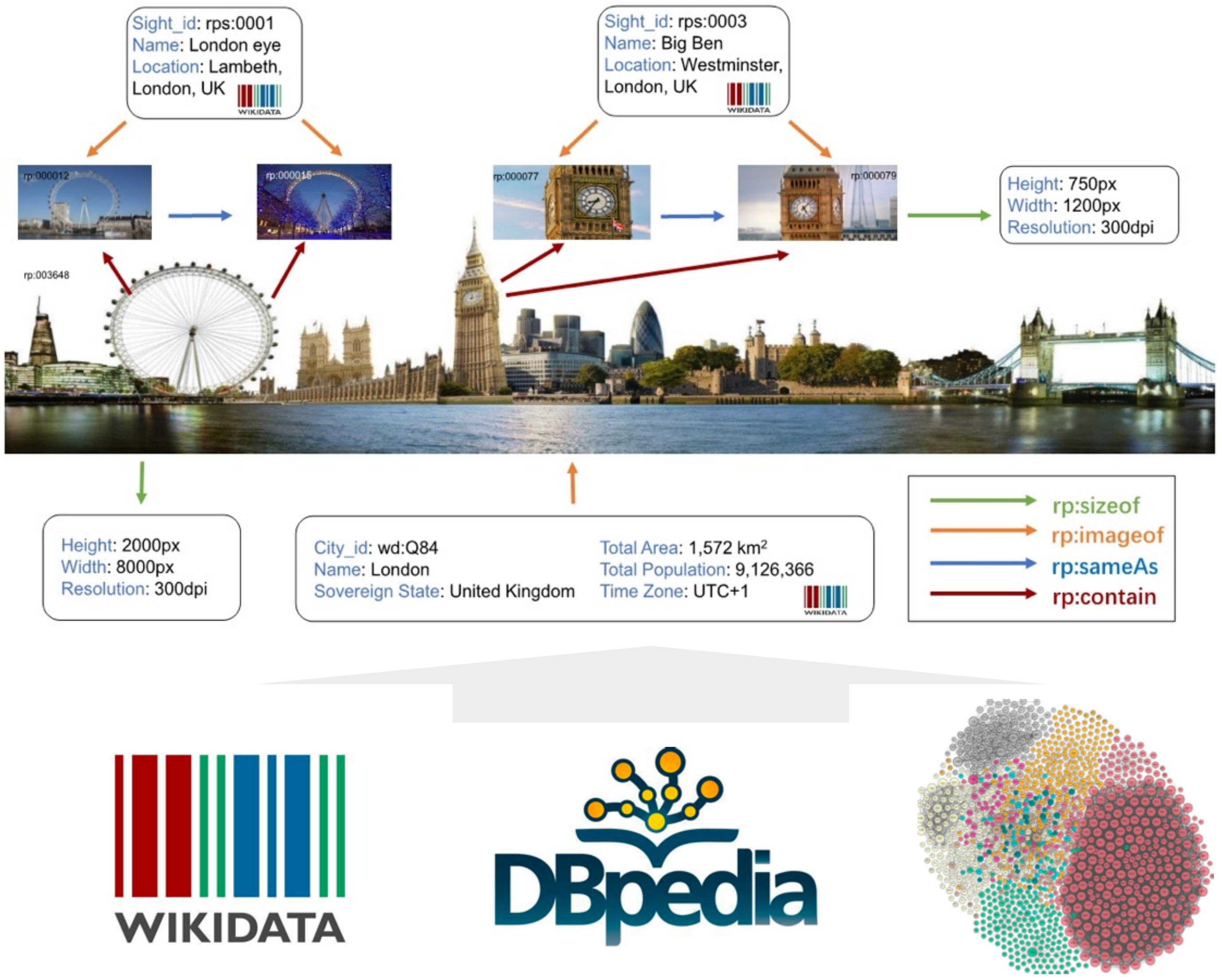


- Multimodality
- Multimodal KG Construction
- Inference
- Challenges


Enhance KG
SPALDING
HOUSTON ROCKETS
Person
stand next
to Person
Man1
(Yao Ming)
beside
Man2
(McGrady)
hold
at
wear
Ball
(Spalding)
Rostrum
(Tian'anm
en)
Jacket
(Houston
Rockets)
Visual relation detection
Cross-modal Entity Linking
Multimodal Knowledge Graph
Application

Enhance KG

1. Enhance KG



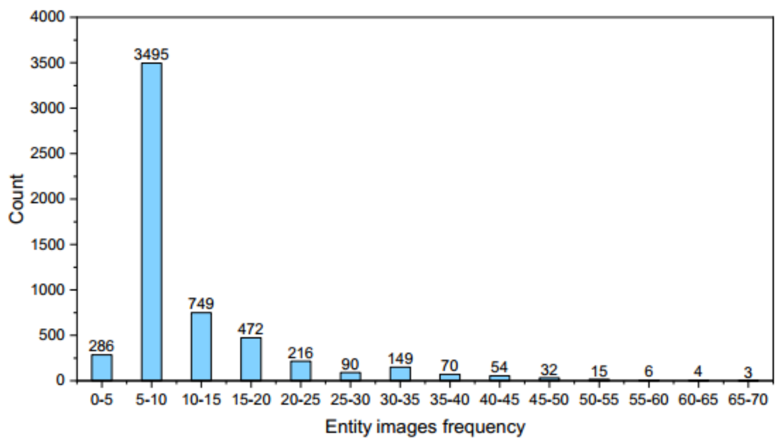
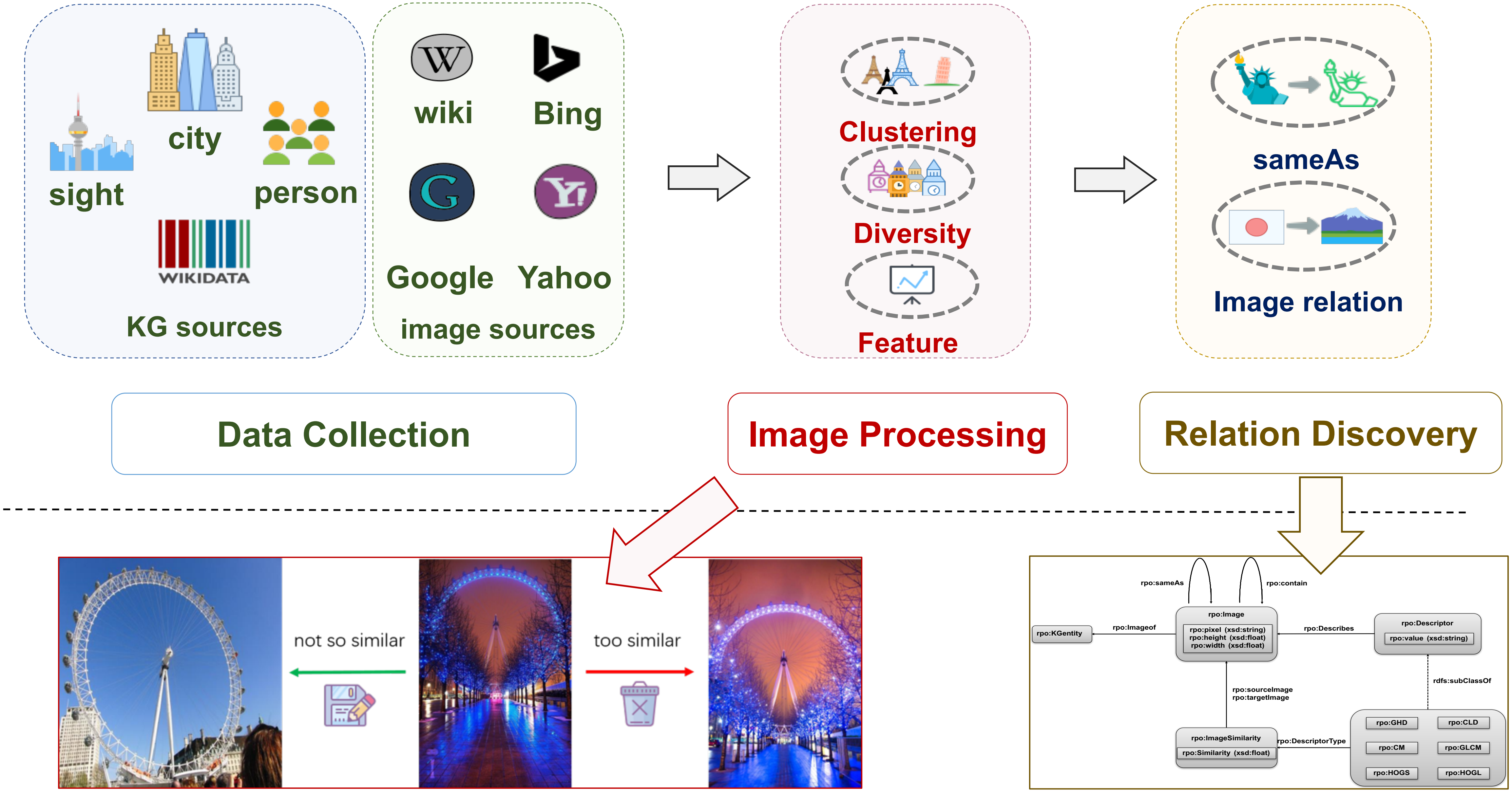


Image in KGs are very sparse

Missing visual relations

1. Enhance KG



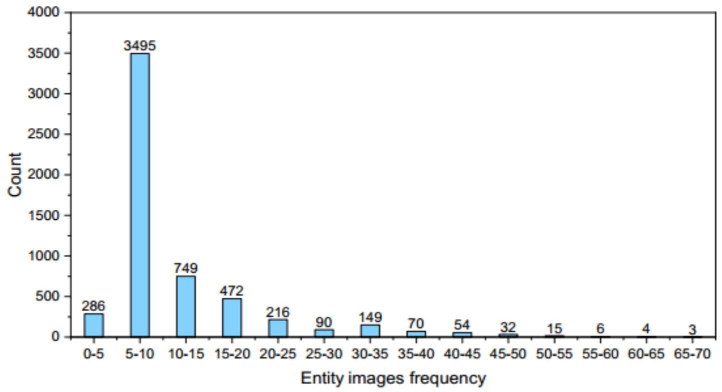
Diversity detection

Visual relation ontology

1. Enhance KG

Richpedia is G, G=(E, R), E: entities R: relations

Motivation: few images



Wikipedia

Bing

Yahoo

Google

Method: collect image resources

- Issue 1: Noise-containing image
- Method: K-means cluster
- Implement: Noise is an outlier in the clusters, and outliers are removed

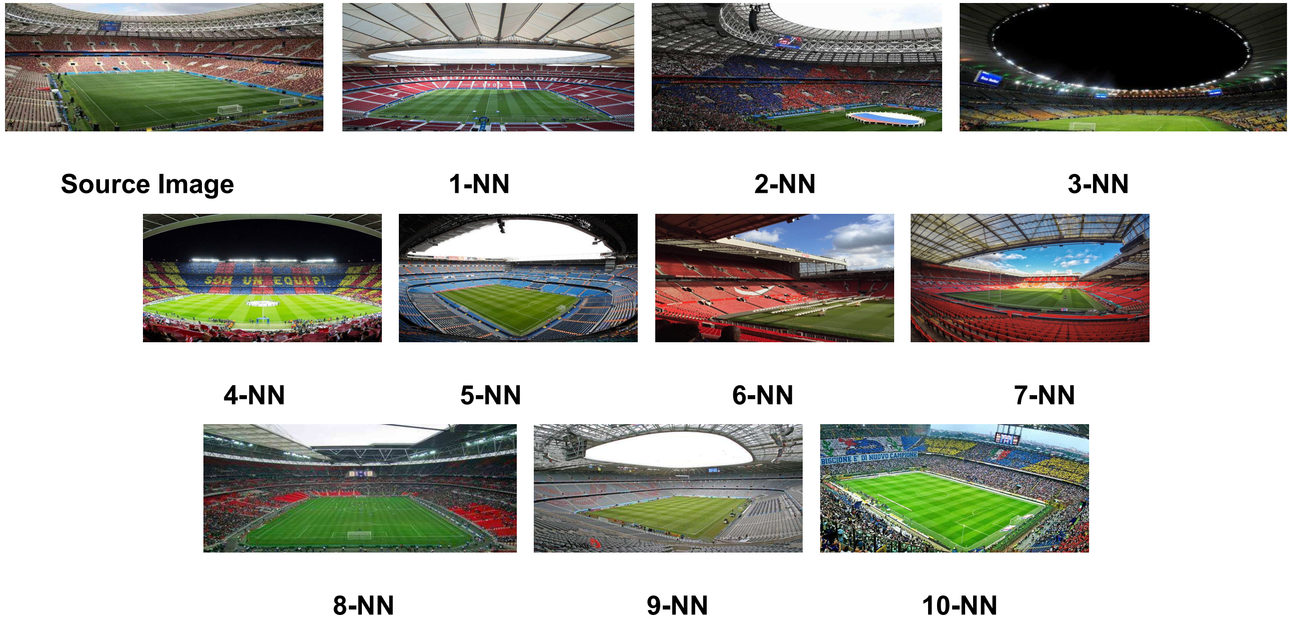
- Issue 2: High image similarity
- Method: Diversity detection algorithm
- Implement:

$$sim(e_i, e_j) = \sum_{k=1}^{n} min(H_k(e_i) - H_k(e_j))$$

First choose the center of mass, then choose the point farthest from the center of mass, and select the farthest point of those selected points in turn.

1. Enhance KG

Richpedia is G, G=(E, R), E: entities R: relations



For each image, we generate five descriptors:

- **rpo:GHD(Gradation Histogram Descriptor)**
- **rpo:CLD(Color Layout Descriptor)**
- **rpo:CMD(Color Moment Descriptor)**
- **rpo:GLCM(Gray-level co-occurrence matrix)**
- **rpo:HOG(Histogram of Oriented Gradient)**

1. Enhance KG

Richpedia is G, G=(E, R), E: entities R: relations

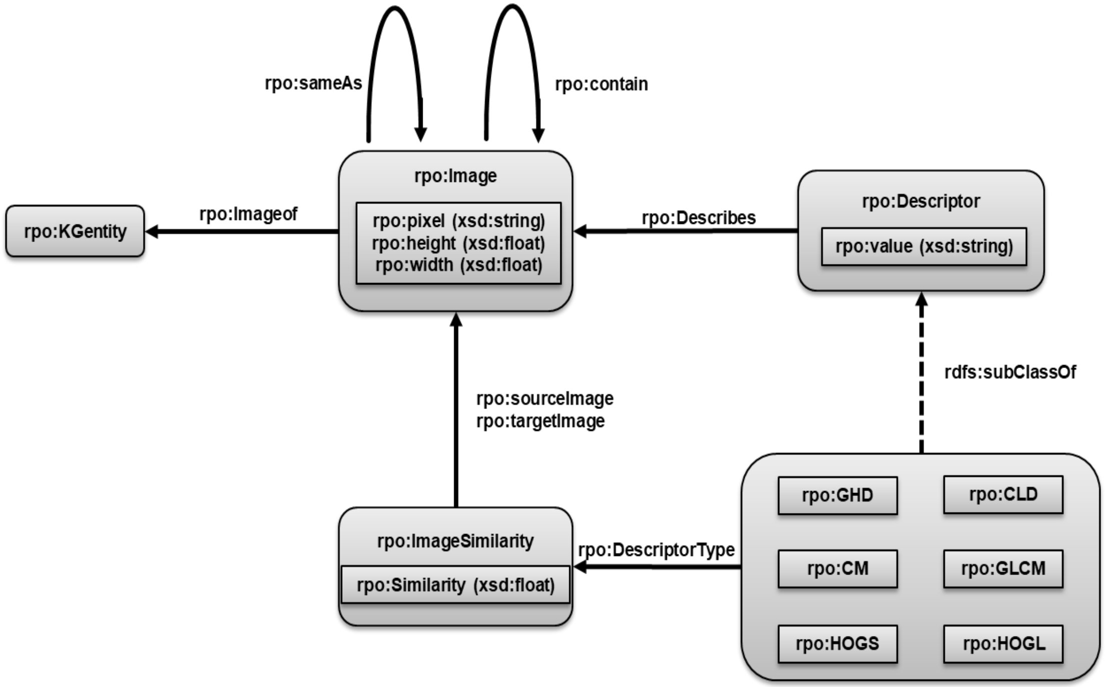
City_id:wd:Q84
Name:London

rpo:imageof

rpo:imageof

rpo:sameAs

<wd:Q84,wdt:P31,wd:Q515>
<rp:0000001,rpo:imageof,wd:Q84>
<rp:0000001,rpo:sameAs,rp:0000002>



Visual relation ontology

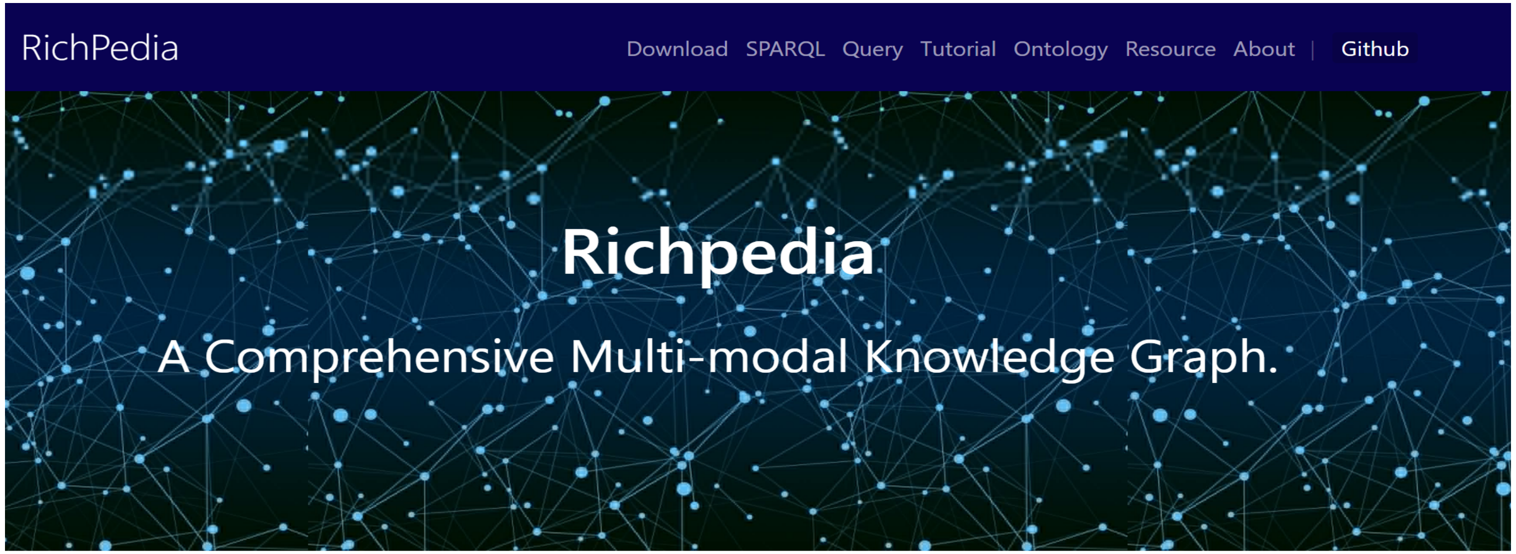
1. Enhance KG

Richpedia is G, G=(E, R), E: entities R: relations

- **Rule based: simple but effective and efficient**

- **Rule1: If there is a hyperlink in the description, the relationship is discovered by a string mapping algorithm between the keyword and the predefined relational ontology.**
- **Rule2: If there are multiple hyperlinks in the description, cyclically input the KG entity corresponding to the hyperlink, simplifying the situation to Rule1.**
- **Rule3: If there are no hyperlinks in the description, use the named entity recognition tool to find the KG entities and simplify the situation to Rule1 and Rule2.**

1. Enhance KG



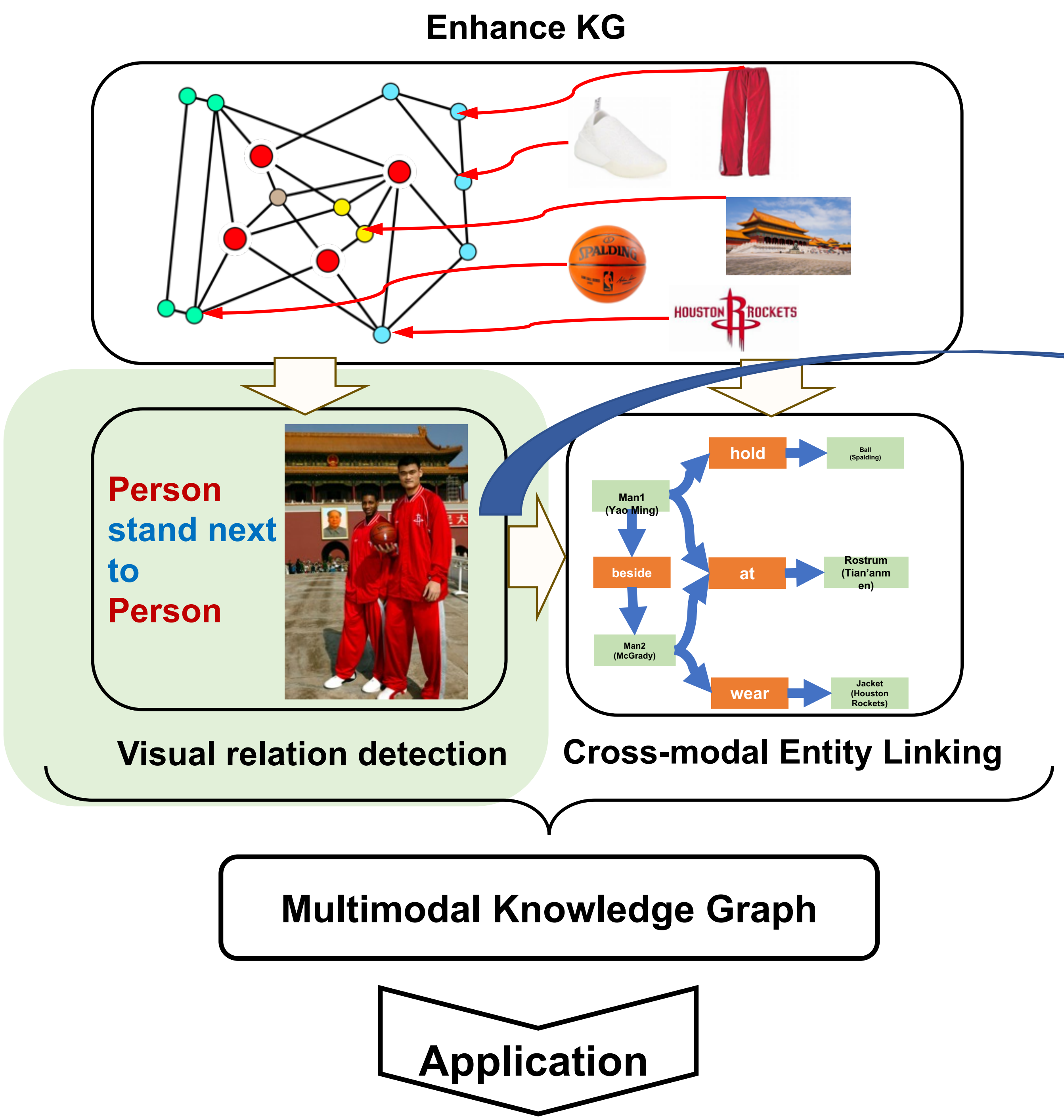
Introduction

With the rapid development of Semantic Web technologies, various knowledge graphs are published on the Web using Resource Description Framework (RDF), such as Wikidata and DBpedia. Knowledge graphs provide for setting RDF links among different entities, thereby forming a large heterogeneous graph, supporting semantic search, question answering and other intelligent services. Meanwhile, public availability of visual resource collections has attracted much attention for different Computer Vision (CV) research purposes, including visual question answering, image classification, object and relationship detection, etc. And we have witnessed promising results by encoding entity and relation information of textual knowledge graphs for CV tasks. Whereas most knowledge graph construction work in the Semantic Web

Richpedia.cn

We will introduce how to use this dataset later

Meng Wang, Guilin Qi, Haofen Wang and Qiushuo Zheng. Richpedia: A Large-Scale, Comprehensive Multi-Modal Knowledge Graph. Big Data Research, 2020



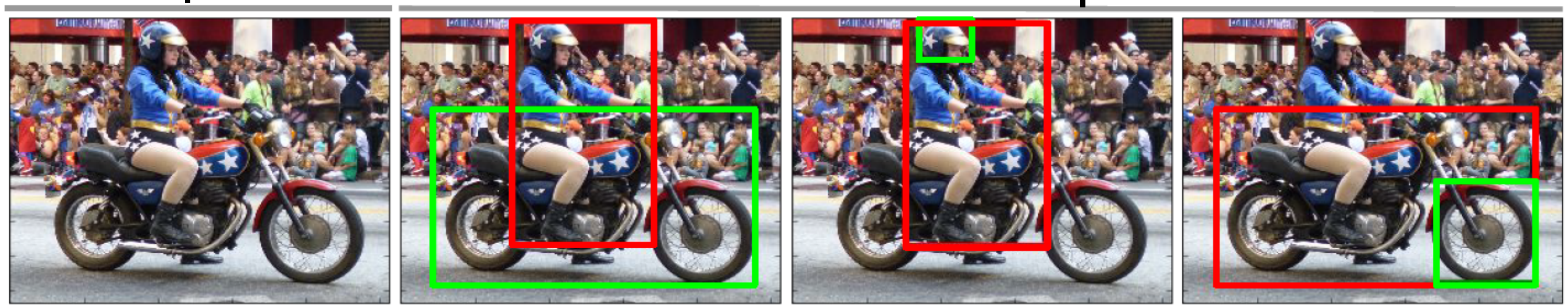
Visual Relation Detection

(Unbalanced, Long-tail)

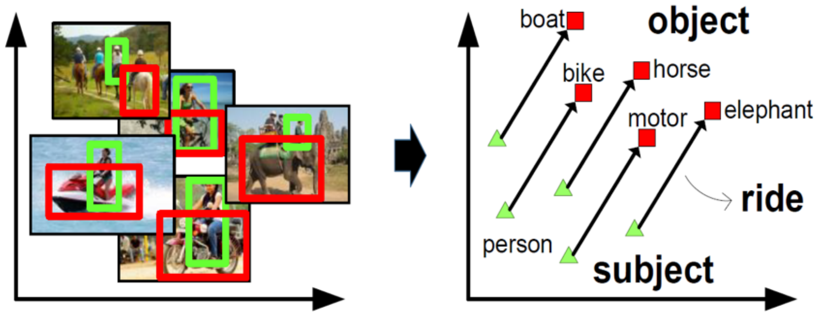
2. Visual Relation Detection

The aim of visual relation detection is to provide a comprehensive understanding of an image by describing all the objects within the scene, and how they relate to each other

Input   Output



person-on-motorcycle   person-wear-helmet   motorcycle-has-wheel

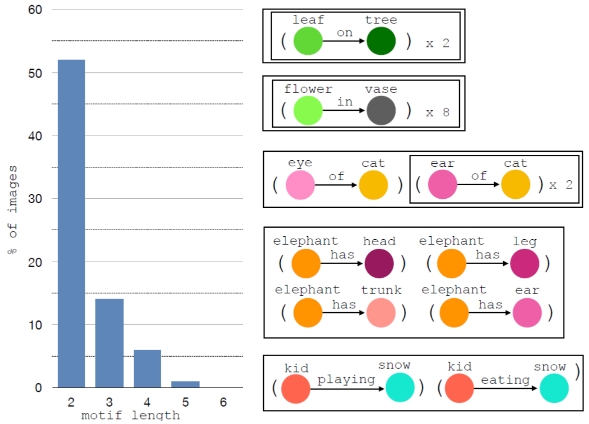


Feature Space   Relation Space

Spatial and Semantic information

VTransE, CVPR 2017

Language Priors, ECCV 2016

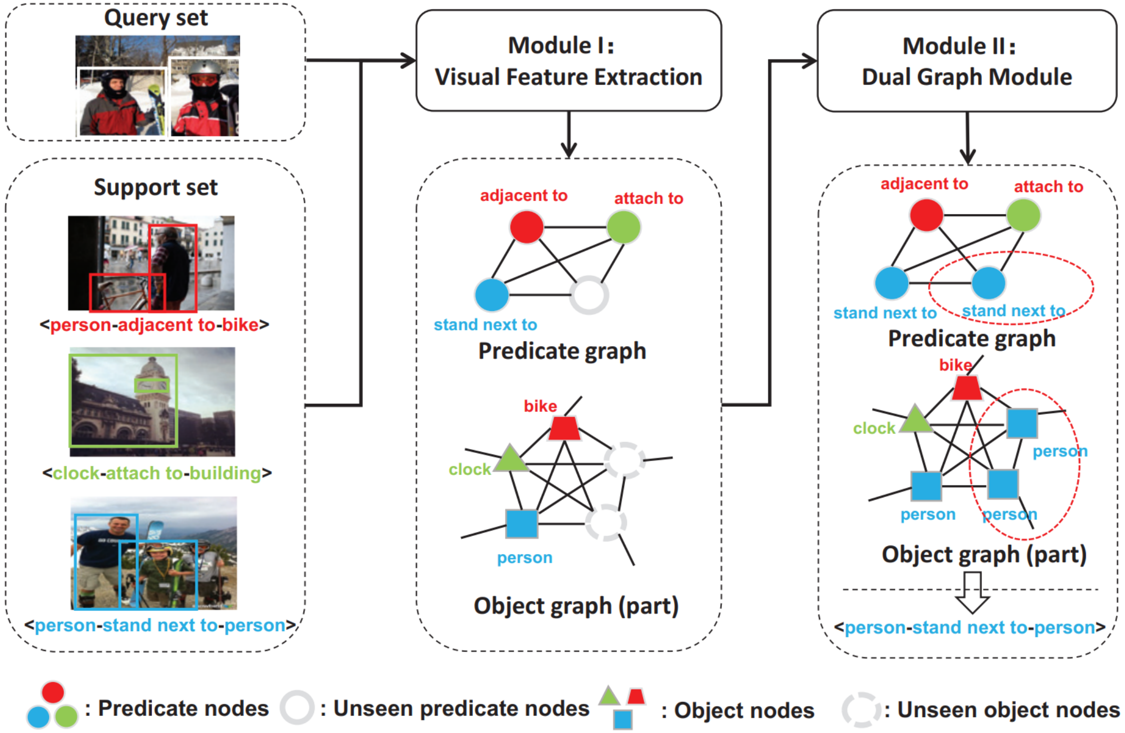


Statistical Information

Neural motifs: Scene graph parsing with global context (CVPR2018)

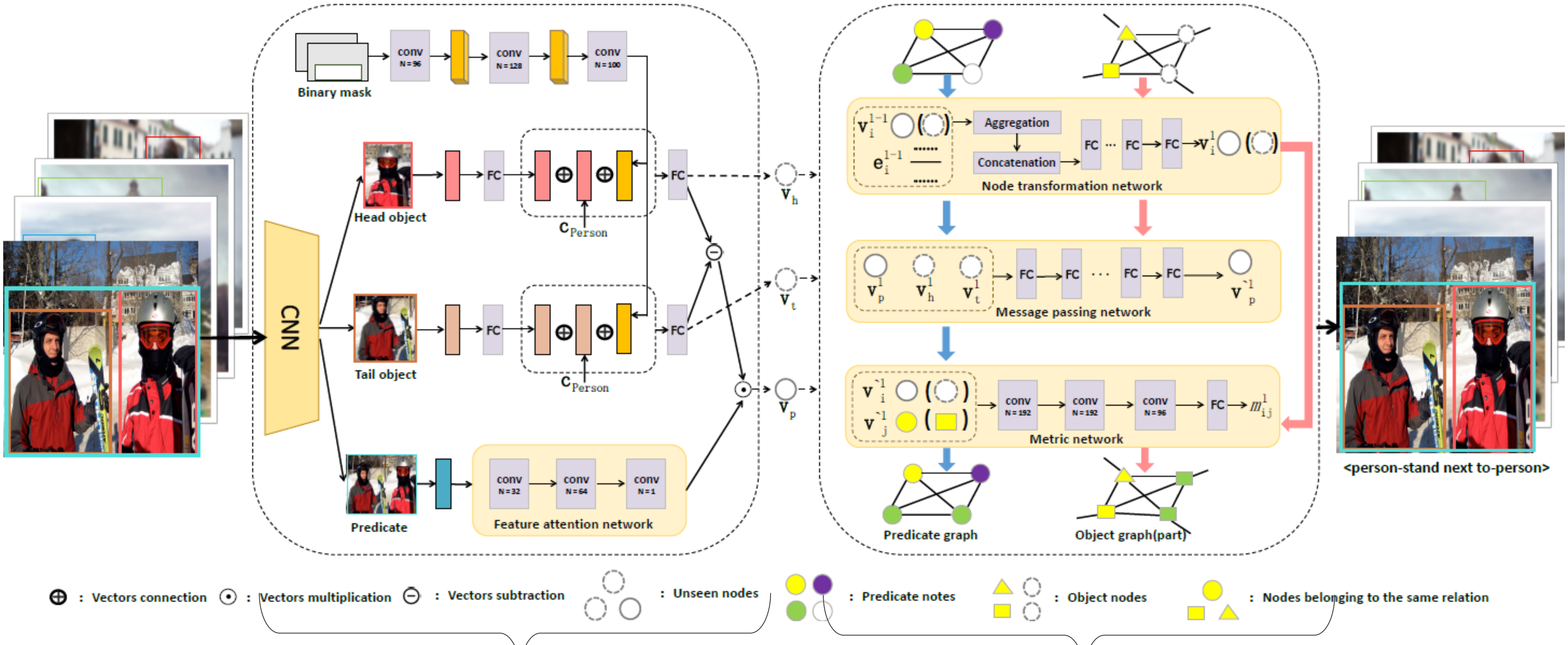
2. Long-tail Visual Relation Detection

Learning to compose dynamic tree structures for visual contexts, CVPR 2019



Our method

2. Long-tail Visual Relation Detection



feature sparsity can be alleviated by using the feature-level attention mechanism

achieve massages passing from predicates or objects in different images

2. Long-tail Visual Relation Detection

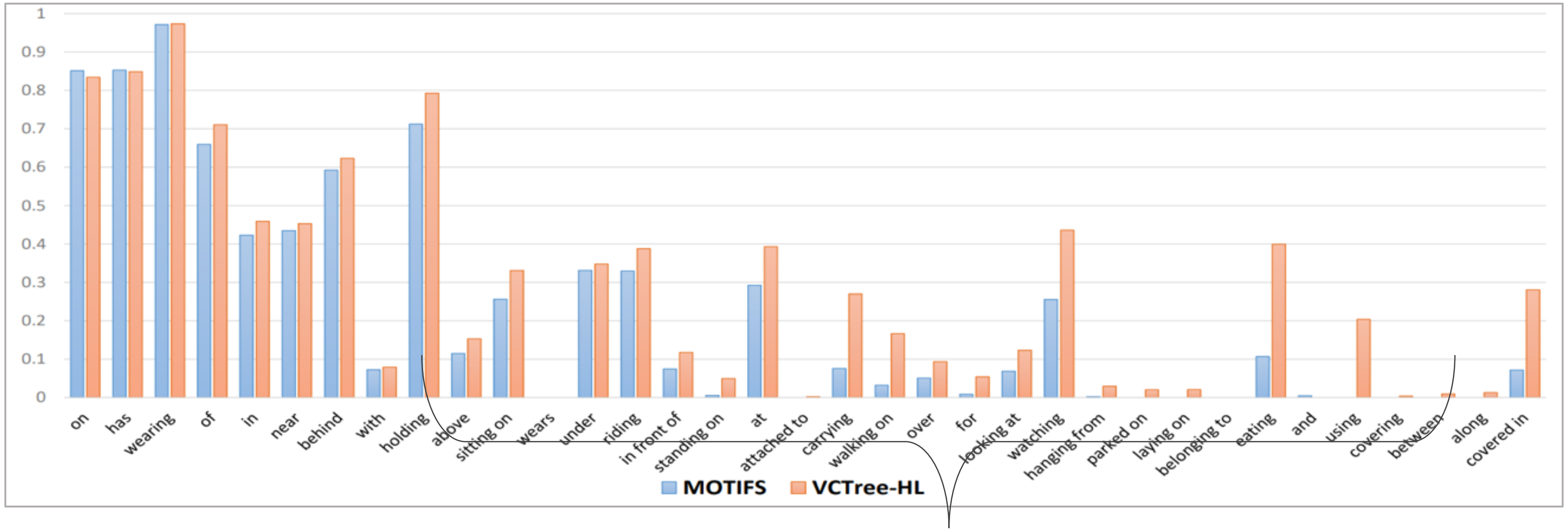


Table 2: Comparison with state-of-the-art baselines on the VRD-One and VG-One datasets.

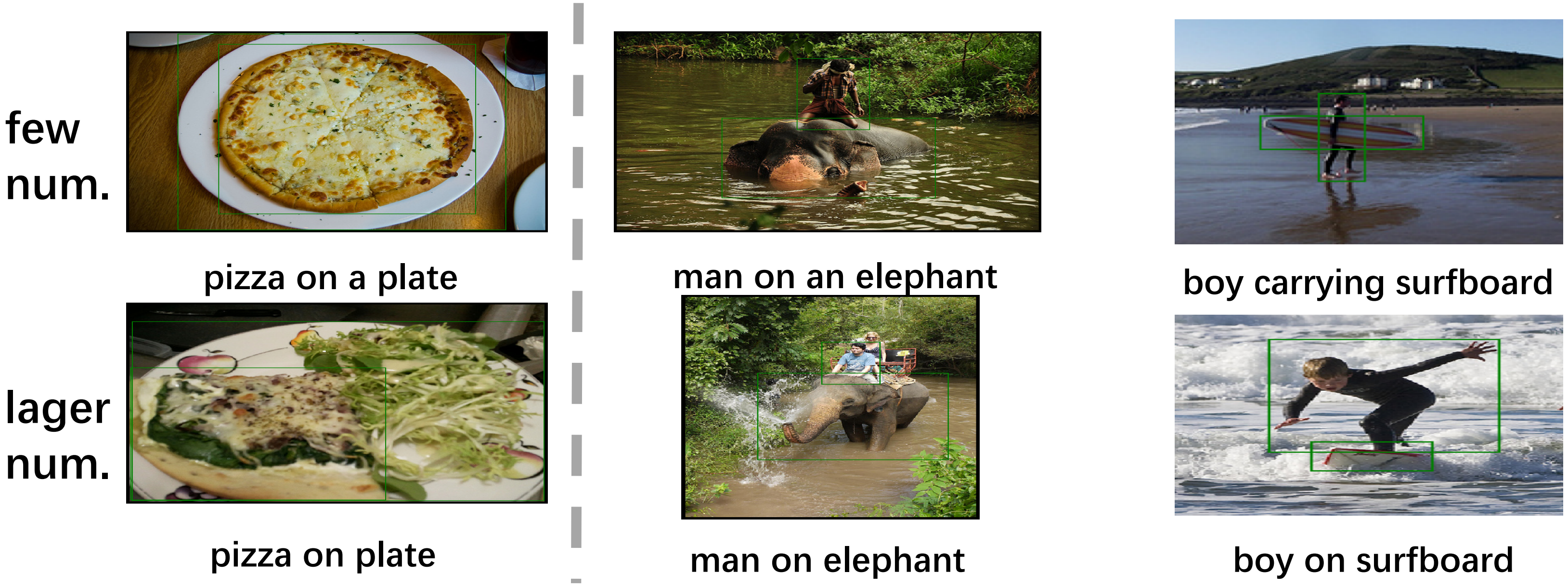
	VRD-One				VG-One			
	PredCls		SGCls		PredCls		SGCls	
	5-way 1-shot	10-way 1-shot	5-way 1-shot	10-way 1-shot	5-way 1-shot	10-way 1-shot	5-way 1-shot	10-way 1-shot
VRD	37.4%	24.7%	14.7%	12.5%	41.4%	27.2%	10.8 %	9.6 %
VTransE	37.3%	24.3%	15.8%	13.4%	39.7%	23.4%	10.1%	9.4 %
LSVRU	40.3%	27.1%	16.9%	14.0%	43.4%	27.0%	10.7%	10.1%
RelDN	40.1%	26.4%	17.2%	14.3%	43.7%	28.3%	11.3%	10.1%
RelDN w/o sem	40.6%	27.3%	17.4%	14.9%	44.1%	28.2%	11.6%	10.4%
Ours	**48.4%**	**33.5%**	**22.3%**	**20.9%**	**56.3%**	**37.5%**	**14.9%**	**13.2%**

Table 4: Ablation studies on our model.

	5-way 1-shot	
	PredCls	SGCls
ours w/o Object Graph	47.3%	20.4%
ours w/o Message Passing Network	47.2%	21.8%
ours w/o Attention Network	45.7%	21.6%
ours All	48.4%	22.3%

Weitao Wang, Meng Wang, Sen Wang, Guodong Long, Lina Yao, and Guilin Qi. One-Shot Learning for Long-Tail Visual Relation Detection. AAAI 2020.

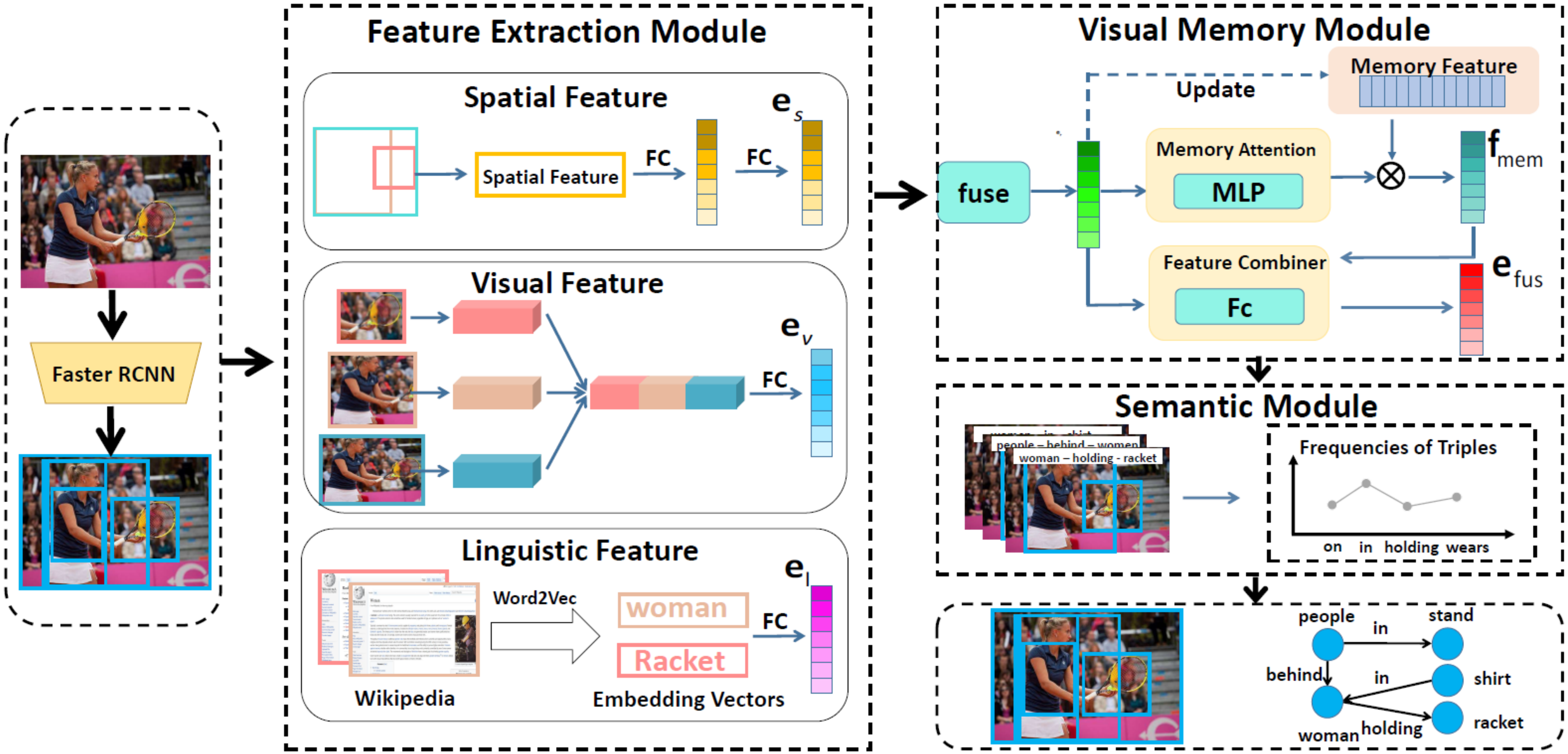
2. Unbalanced Visual Relation Detection



(a) Nonstandard label (b) Features overlap

Due to the existence of nonstandard labels, excessive attention to low-frequency visual relation will affect the performance of the scene graph generation model.

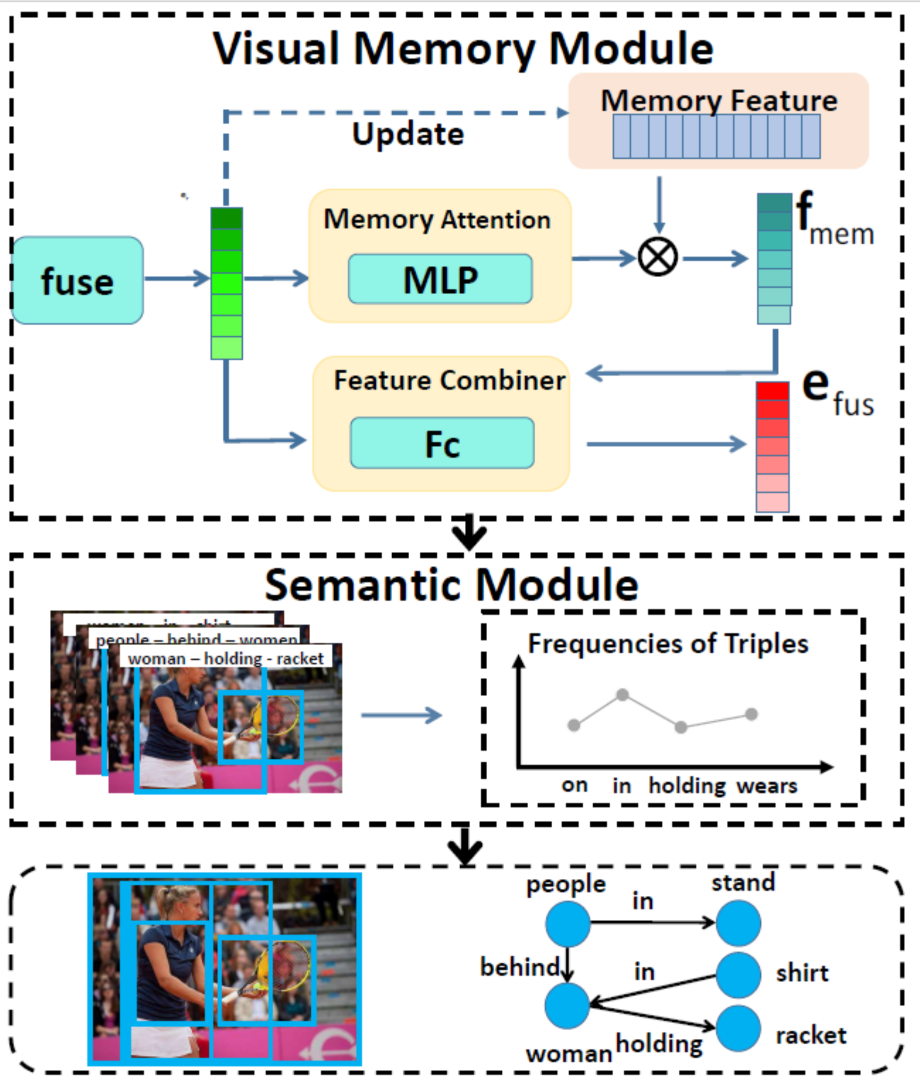
2. Unbalanced Visual Relation Detection



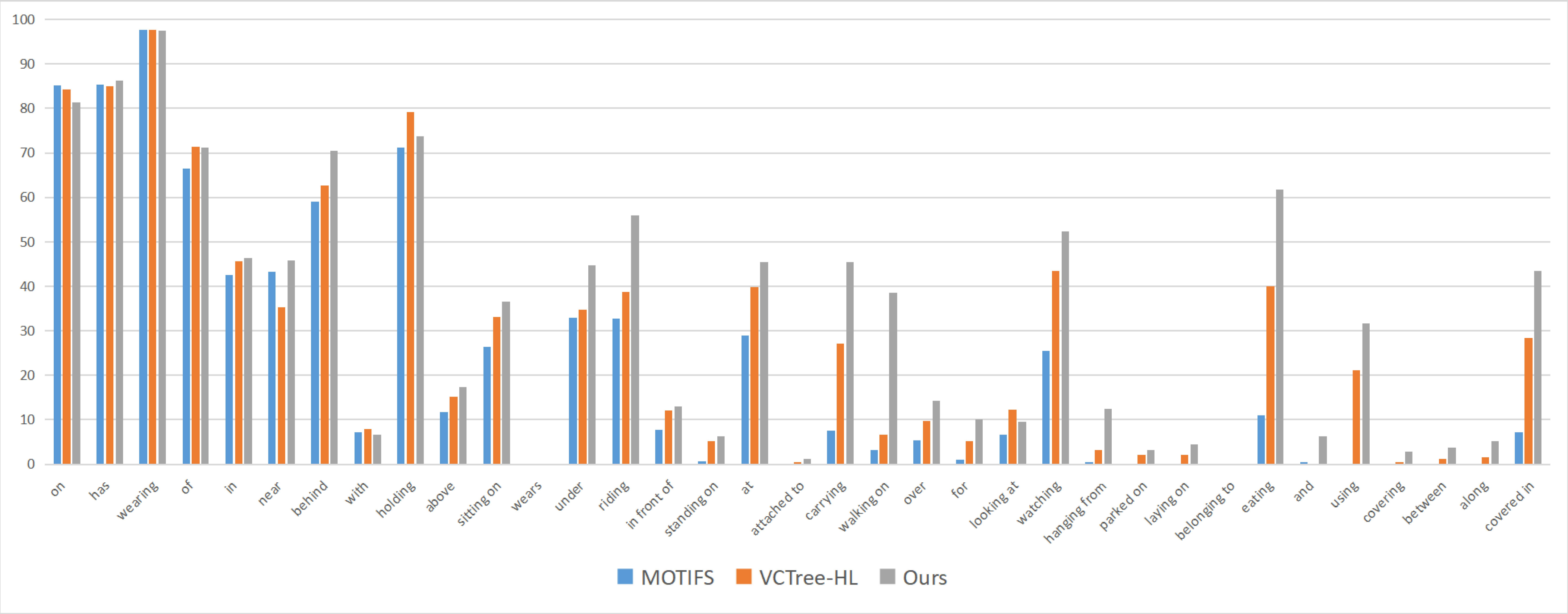
We use the method of memory features to realize the transfer of high-frequency relation features to low-frequency relation features.

2. Unbalanced Visual Relation Detection

- **The calculation of visual relation memory is based on the prototype of each class, which is the mean of each category of features in the training set.**
- **The direct observation features and memory features are fused to realize the information exchange between the current relation and other relations**
- **We also utilize the statistical information (distribution) from the training set to influence the results of the model.**

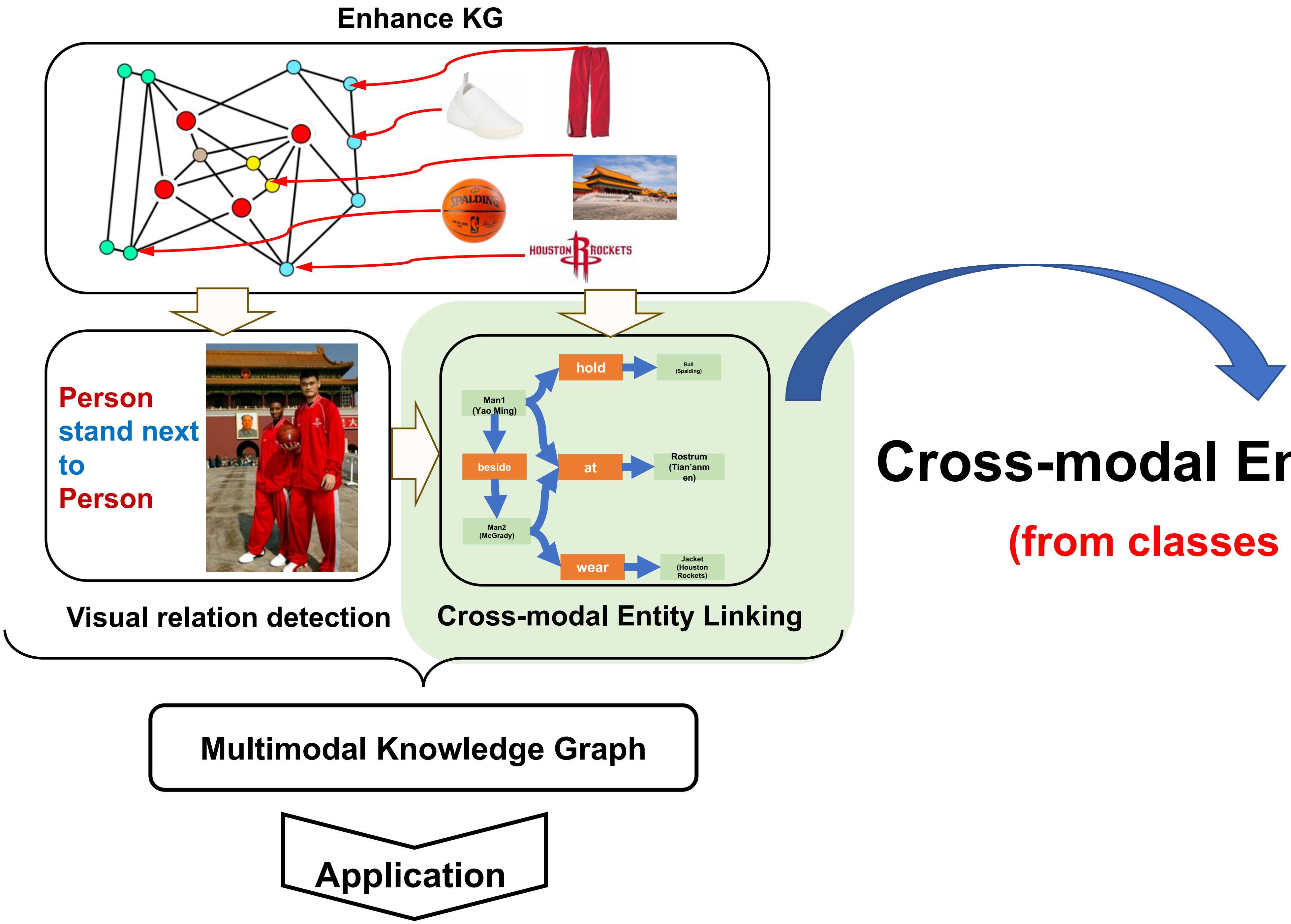


2. Unbalanced Visual Relation Detection



Our model achieves evident improvement in almost all relations (47/50)

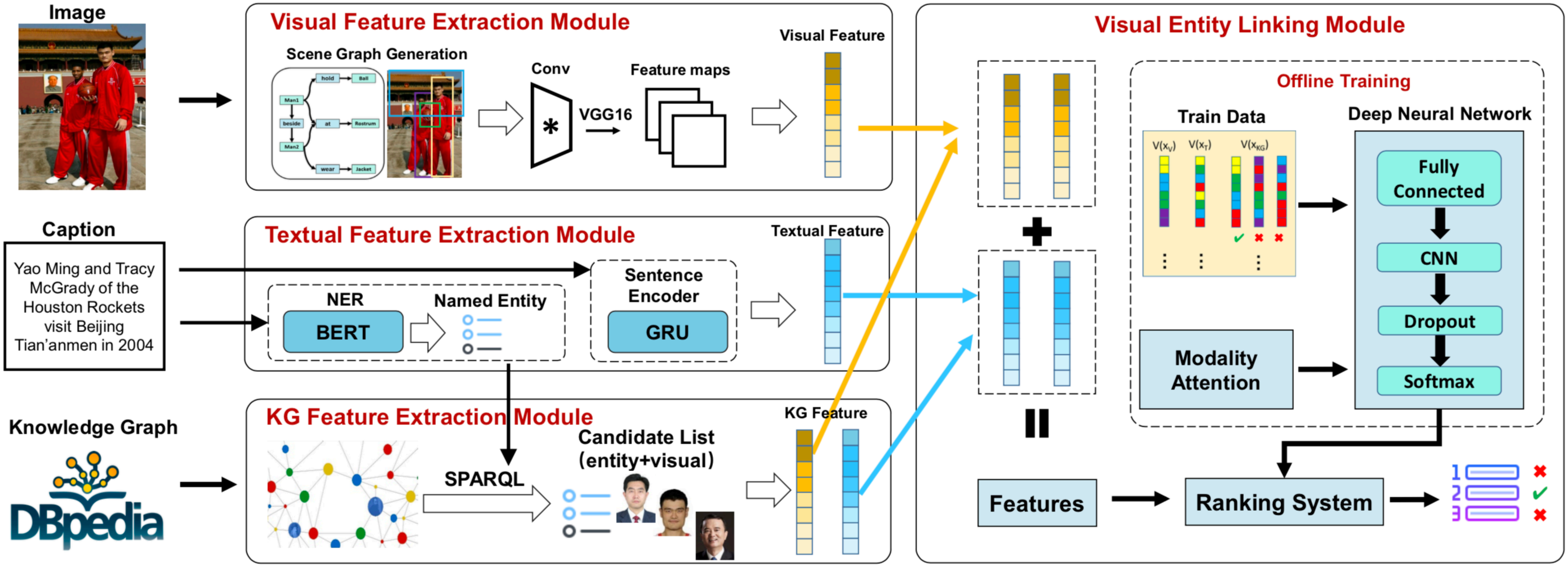
Weitao Wang, Ruyang Liu, Meng Wang, Sen Wang, Xiaojun Chang, and Yang Chen. Memory-Based Network for Scene Graph with Unbalanced Relations. 28th ACM MM 2020.



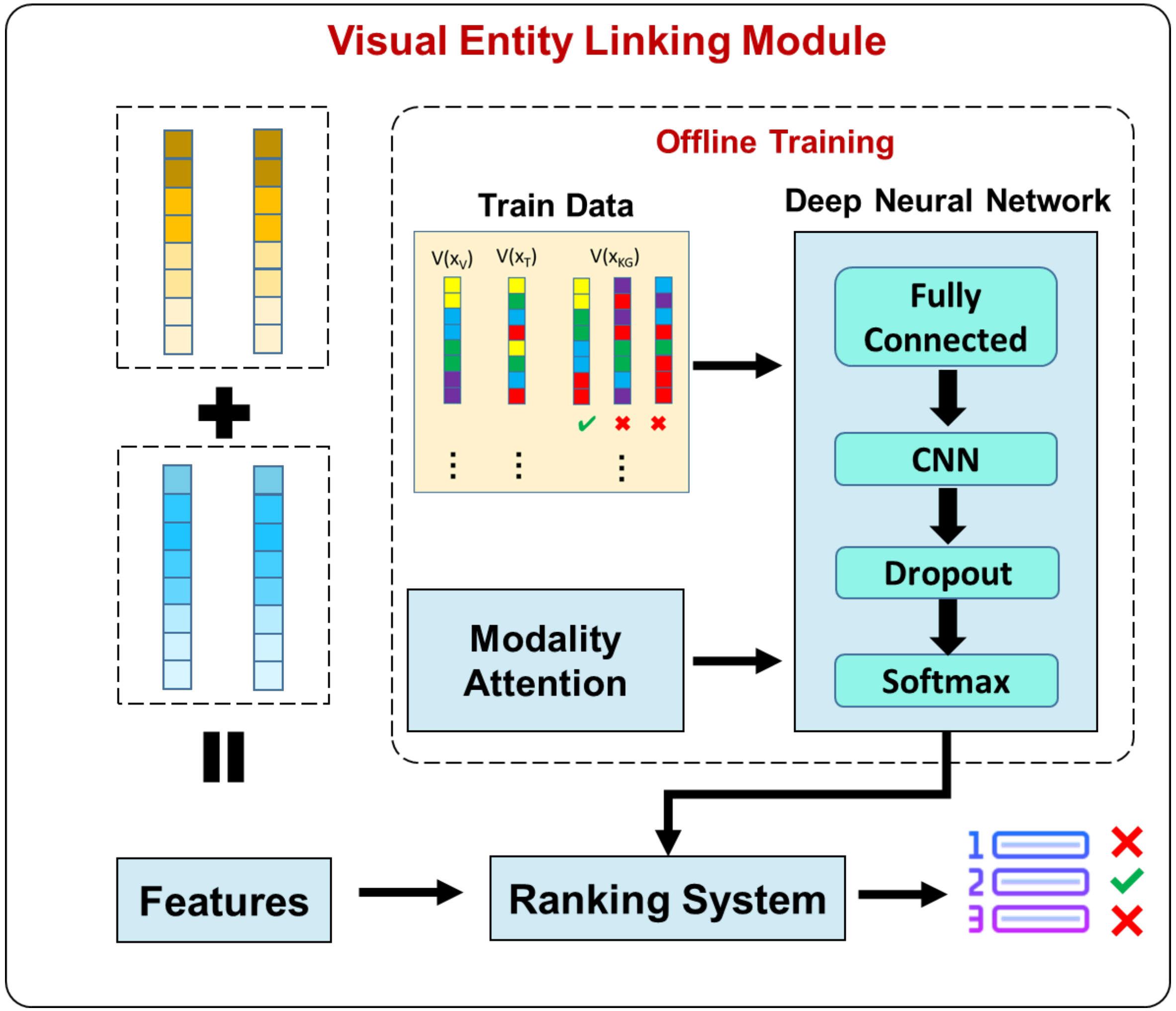
Cross-modal Entity Linking

(from classes to entities)

3. Cross-modal Entity Linking



3. Cross-modal Entity Linking



$$\mathcal{L} = \sum_{s=1}^{m} \left(\lambda_t \mathcal{L}_{\mathrm{T}}\left(\bar{\mathbf{x}}^s\right) + \lambda_v \mathcal{L}_{\mathrm{V}}\left(\bar{\mathbf{x}}^s\right)\right) + \|\mathbb{W}\|_2^2$$

$$\mathcal{L}_{\mathrm{T}}\left(\bar{\mathbf{x}}\right) = \sum \left[\gamma + conf_t\left(\mathbf{y}'\right) - conf_t\left(\mathbf{y}\right)\right]_+$$

$$\mathcal{L}_{\mathrm{V}}\left(\bar{\mathbf{x}}\right) = \sum \left[\gamma + conf_v\left(\mathbf{y}'\right) - conf_v\left(\mathbf{y}\right)\right]_+$$

$$conf_t(\mathbf{y}^i) = \frac{\exp(f(\mathbf{e}\left(\mathbf{x}_t\right), \mathbf{e}\left(\mathbf{y}_t^i\right)))}{\sum_{j\in\mathbb{C}} \exp(f(\mathbf{e}\left(\mathbf{x}_t\right), \mathbf{e}\left(\mathbf{y}_t^j\right)))}$$

$$conf_v(\mathbf{y}^i) = cosine\left(\mathbf{e}\left(\mathbf{x}_v\right), \mathbf{e}\left(\mathbf{y}_v^i\right)\right)$$

Learning to Rank

With

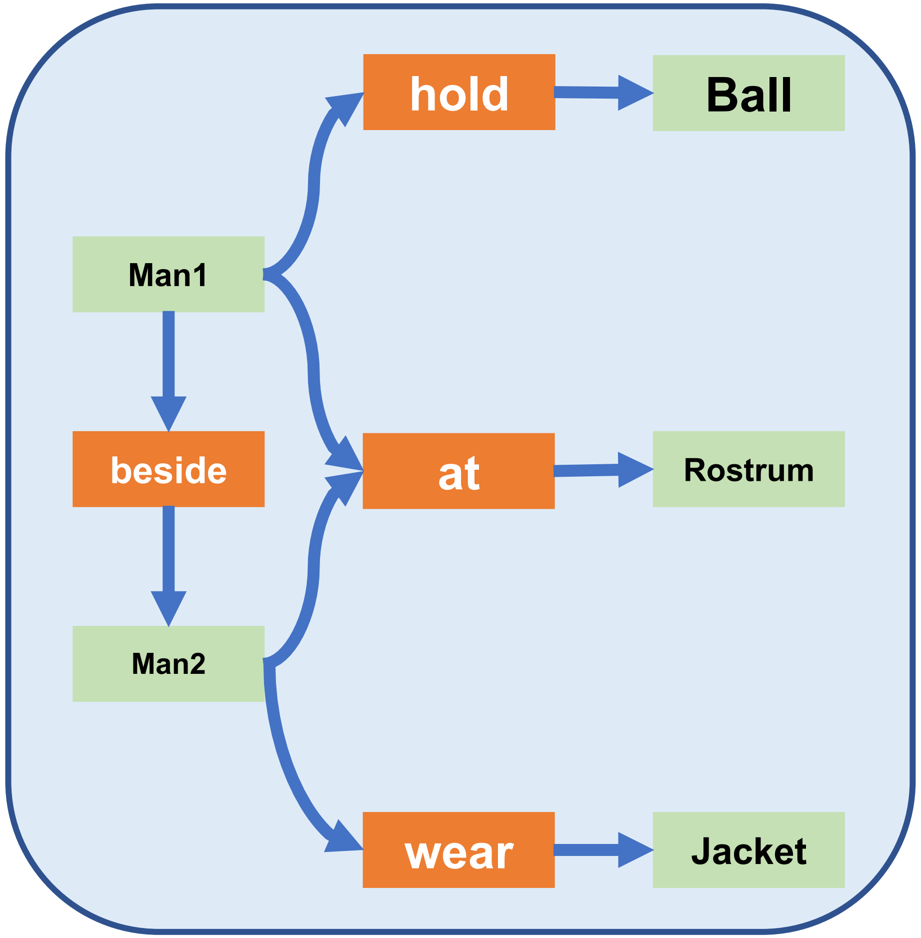
Modal Attention

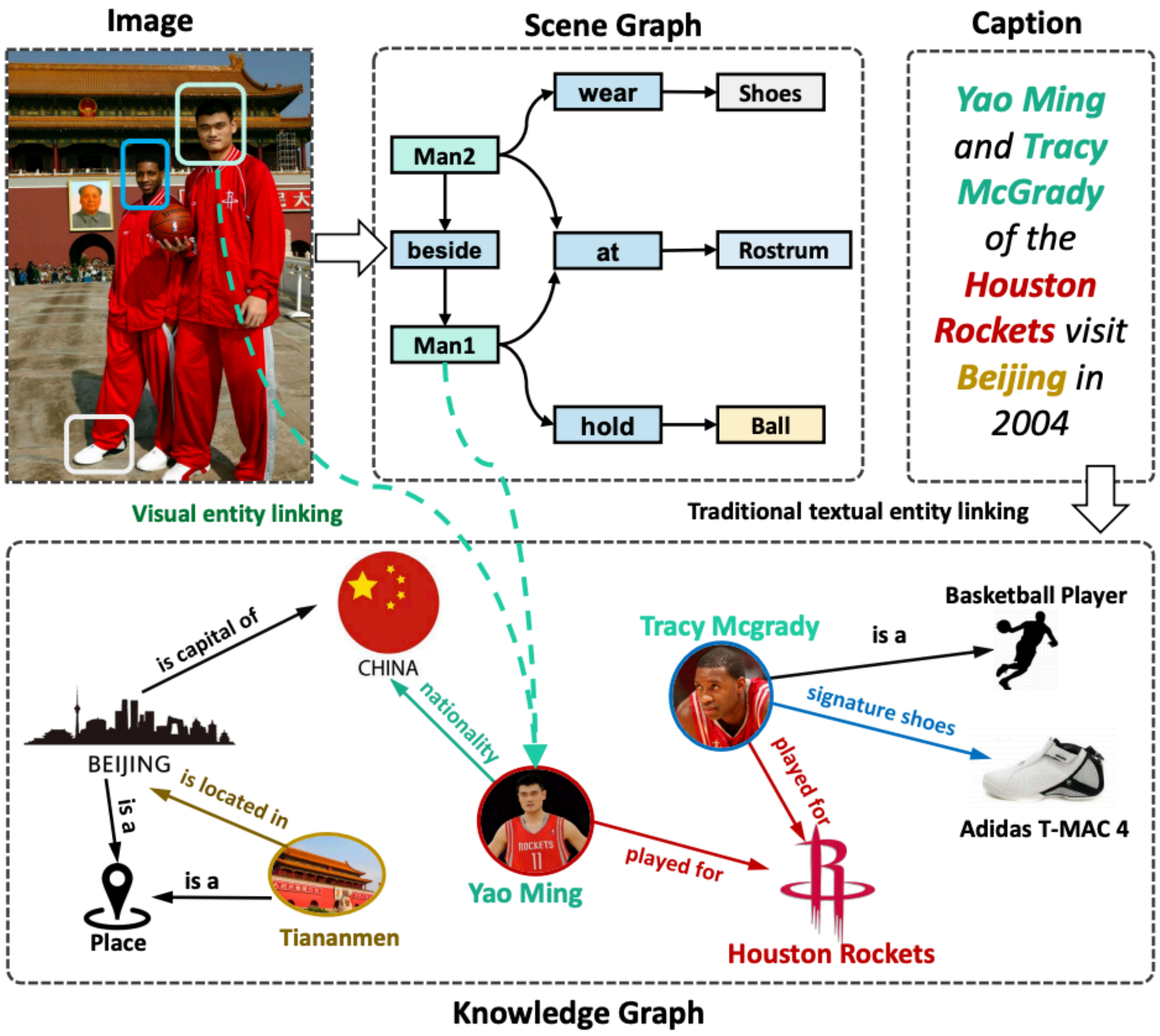
$$\left[\mathbf{a}_t; \mathbf{a}_v\right] = \sigma\left(\mathbf{W} \cdot \left[\mathbf{x}_t; \mathbf{x}_v\right] + \mathbf{b}\right)$$

$$\alpha_m = \frac{\exp\left(\mathbf{a}_m\right)}{\sum_{m'\in\{t,v\}} \exp\left(\mathbf{a}_{m'}\right)} \quad \forall m \in \{t, v\}$$

$$\bar{\mathbf{x}} = \sum_{m\in\{t,v\}} \alpha_m \mathbf{x}_m$$

3. Cross-modal Entity Linking

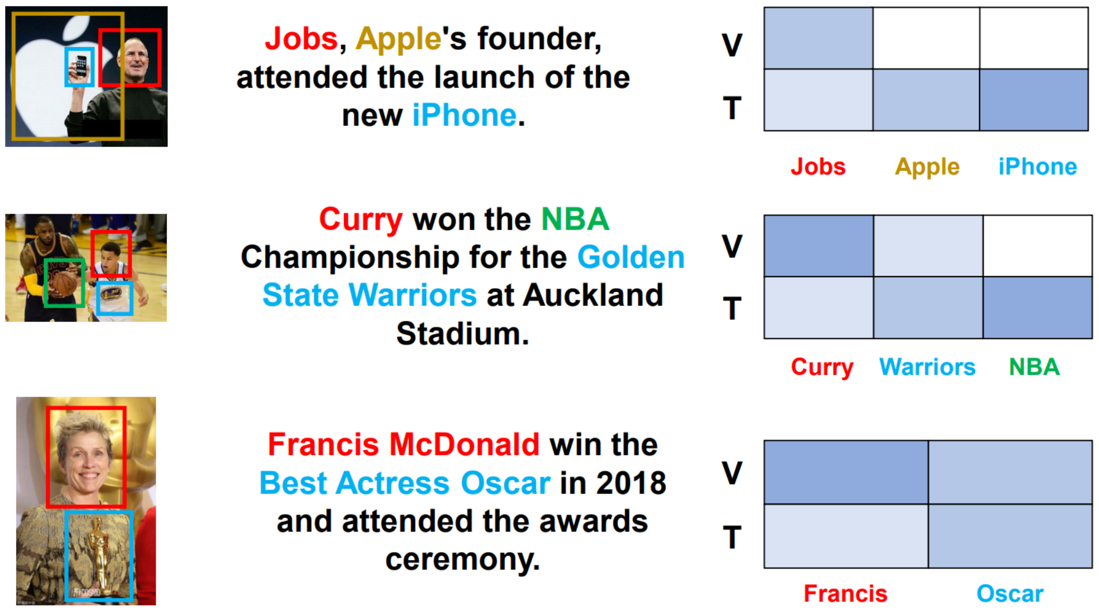


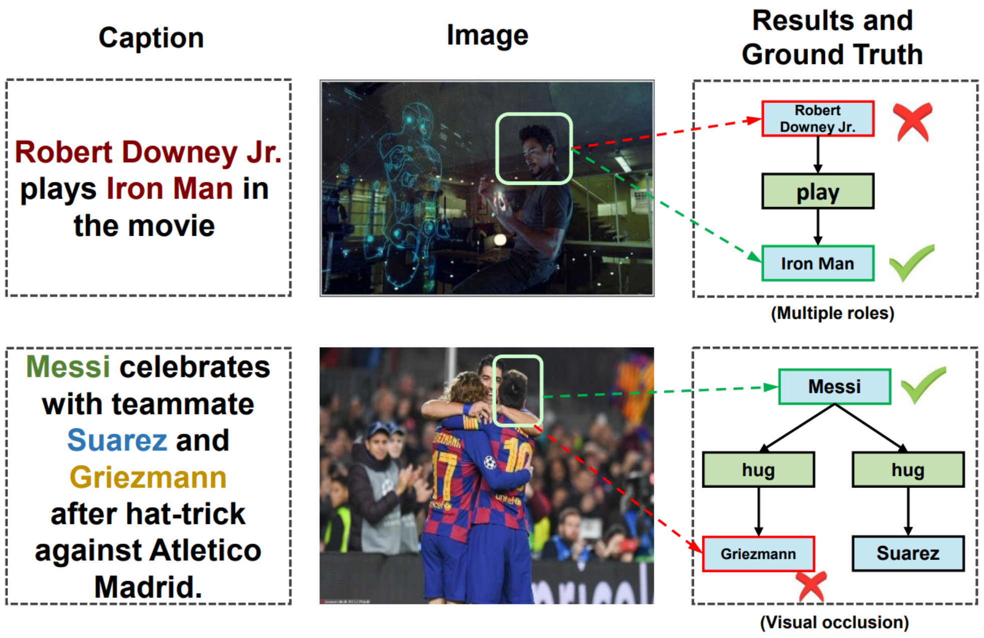


3. Cross-modal Entity Linking

	Accuracy	
	Top-1	Top-10
ours w/o visual features	56.19%	63.42%
ours w/o textual features	72.55%	82.23%
ours with a smaller sized KG	60.19%	66.47%
ours All	**83.16%**	**93.81%**

Modalities	Model	Accuracy			
		Top-1	Top-3	Top-5	Top-10
V+KG	Huawei API	12.53%	18.49%	20.46%	22.94%
V+KG	Tencent API	11.79%	16.42%	21.64%	24.61%
T+KG	Faster-RCNN+CoAtt	55.45%	63.76%	66.05%	67.91%
T+KG	Faster-RCNN+Falcon	56.16%	61.47%	62.17%	63.94%
T+KG	Faster-RCNN+CDTE	58.27%	64.79%	65.09%	66.14%
V+T+KG	DZMNED	66.46%	73.16%	81.06%	83.49%
V+T+KG	**Our method**	**83.16%**	**88.61%**	**92.49%**	**93.81%**

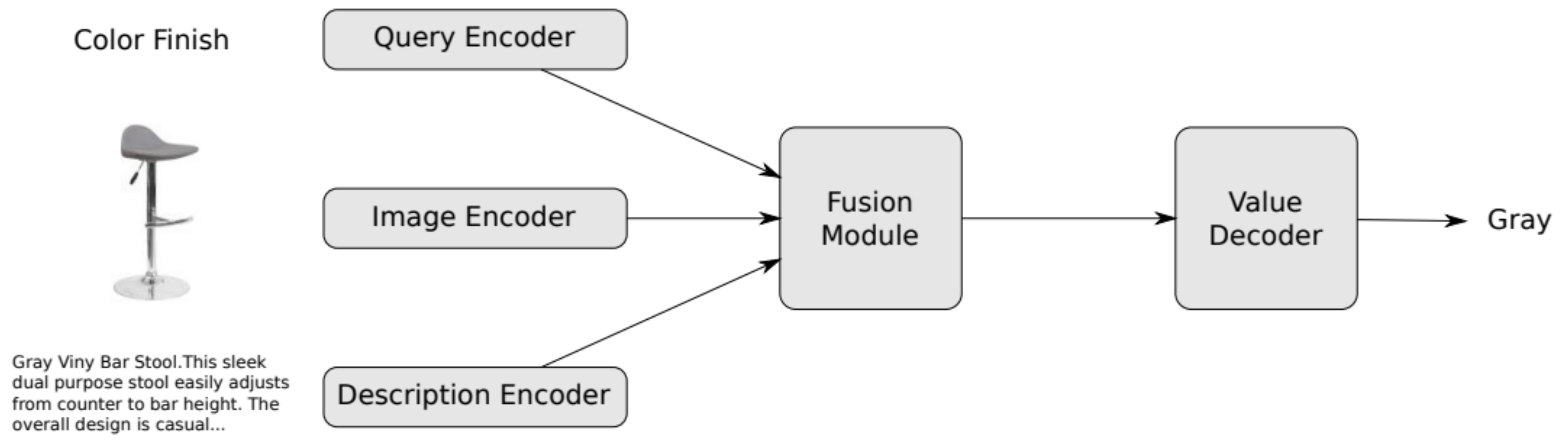


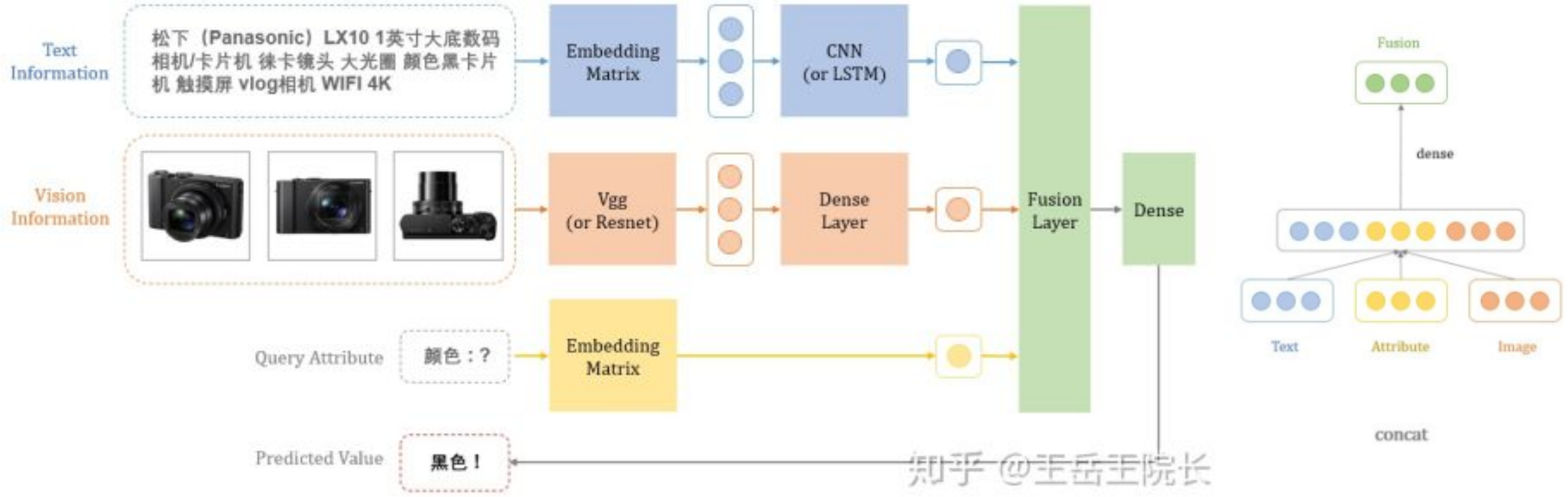


Qiushuo Zheng, Meng Wang, Guilin Qi, Chaoyu Bai. Semantic Visual Entity Linking based on Multimodal Learning. *AAAI*. 2021 (submitted)

- Multimodality
- Multimodal KG Construction
- Inference
- Challenges

Multi-modal KG Completion





Logan IV R L, Humeau S, Singh S. Multimodal attribute extraction. NIPS, 2017.

Multi-modal KG Completion

Input

+ Item Details

Output

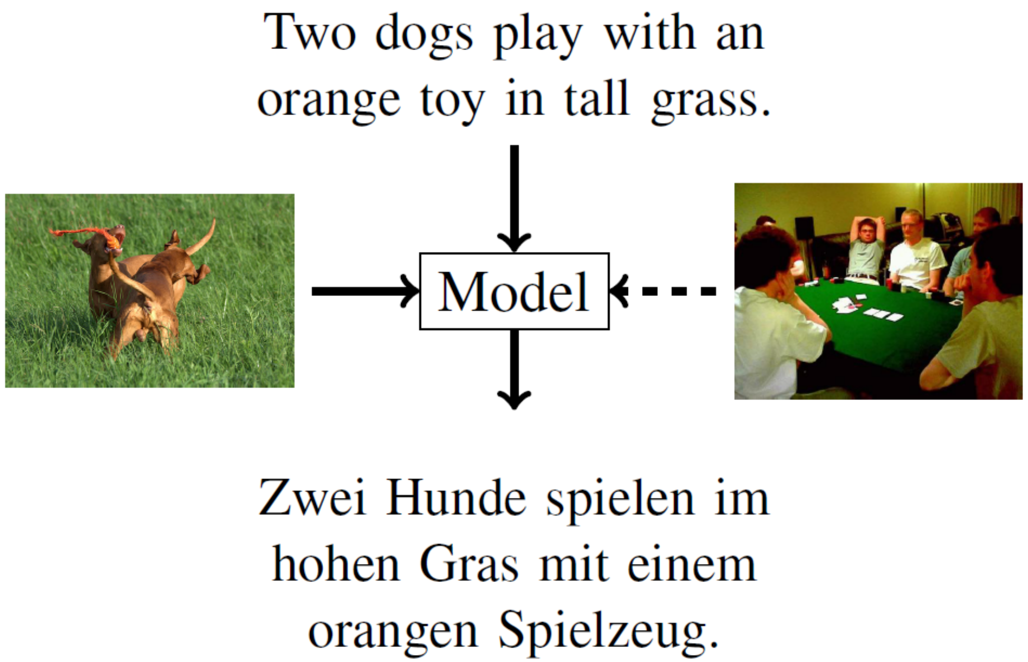
商品名称：芝华仕沙发 | 商品编号：100000993577 | 商品毛重：116.0kg | 商品产地：重庆、苏州、惠州、...
货号：8908A | 功能类别：功能沙发 | 材质类别：布艺沙发 | 适用面积：中户型（60㎡-90㎡）
适用人数： | 填充物：海绵 | 沙发组合样式：一字型 | 是否可定制：不可定制
面料材质：科技布 | 是否带储物：不可储物 | 主色系： | 是否带贵妃榻：
风格：现代简约 | 是否可拆洗：不可拆洗 | 框架材质：板木结合 | 适用场景：客厅，庭院，阳台

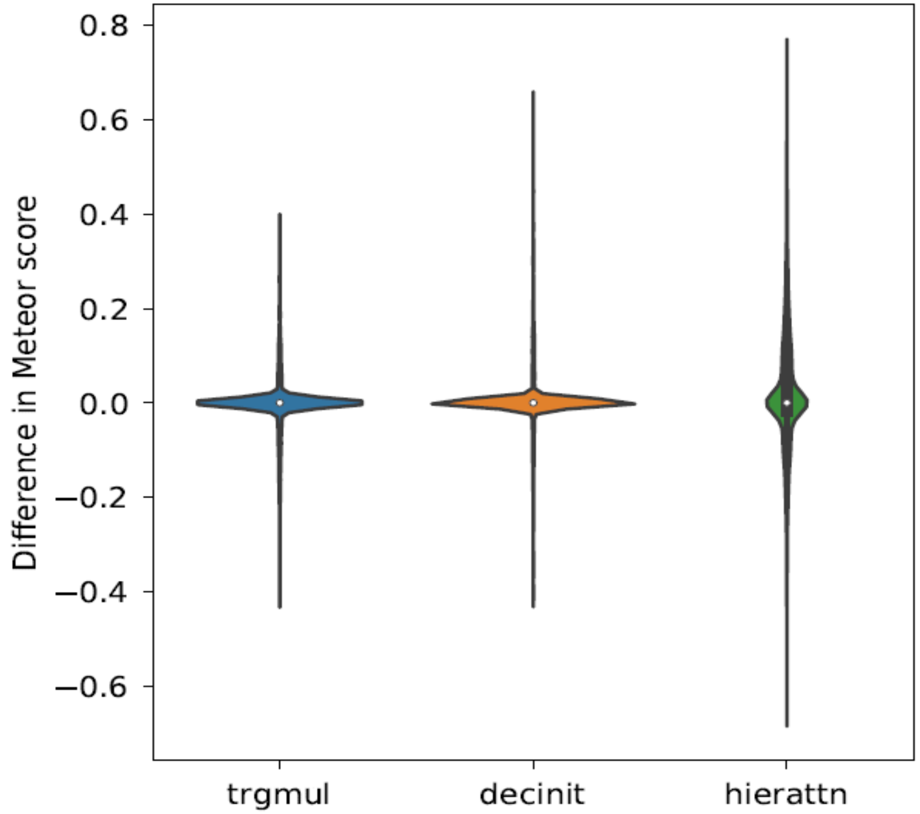
[三人、白色/浅色、不带]

Our MKG base model

Is the multi-modal KG really helpful?

Adversarial Evaluation of Multimodal Machine Translation. EMNLP 2018. (CCF B)

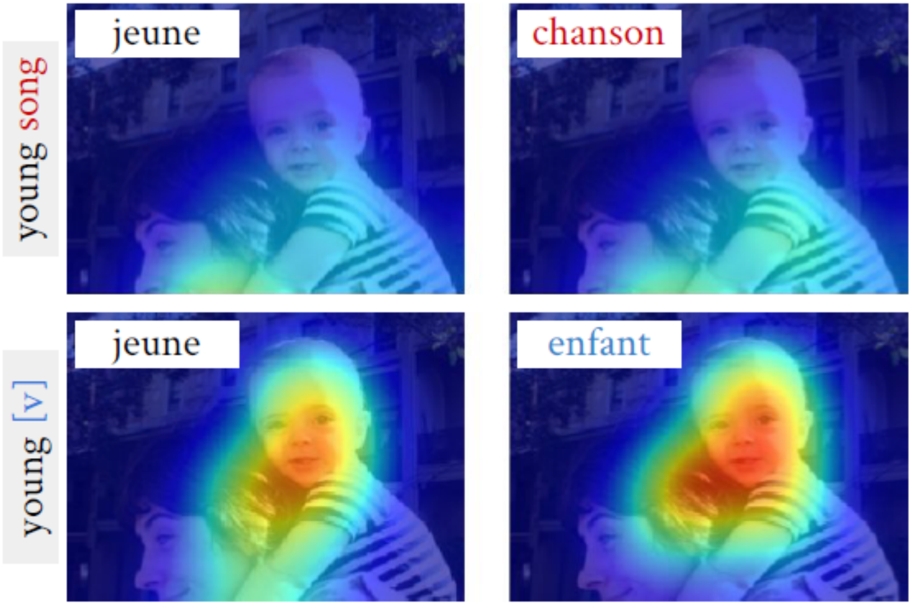


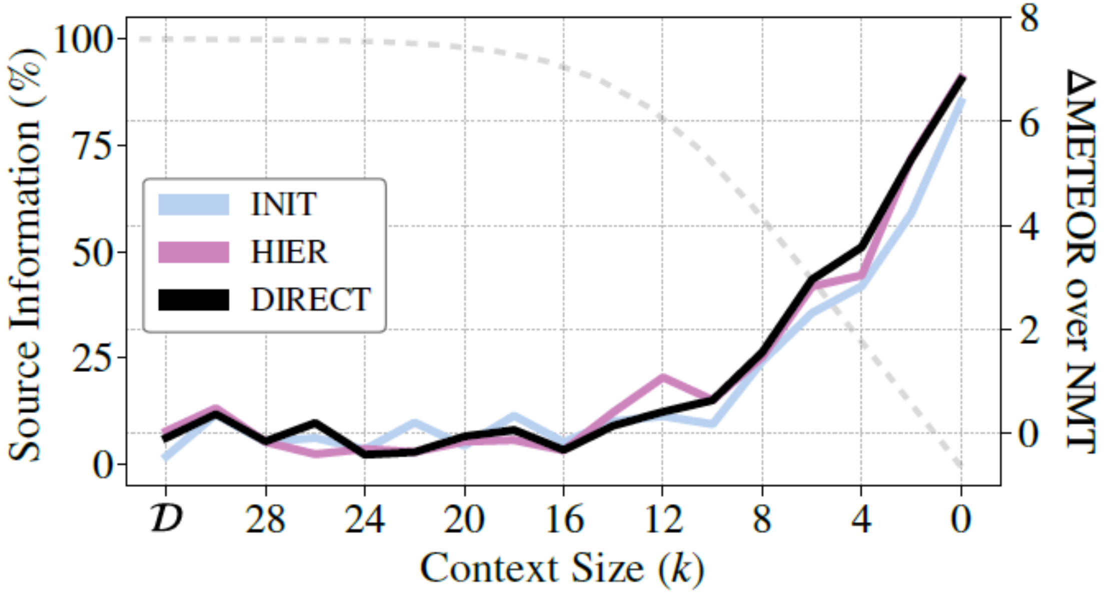


Only needed for incorrect, ambiguous, and gender-neutral words

Is the multi-modal KG really helpful?

Probing the Need for Visual Context in Multimodal Machine Translation.
NAACL 2019. (CCF B)

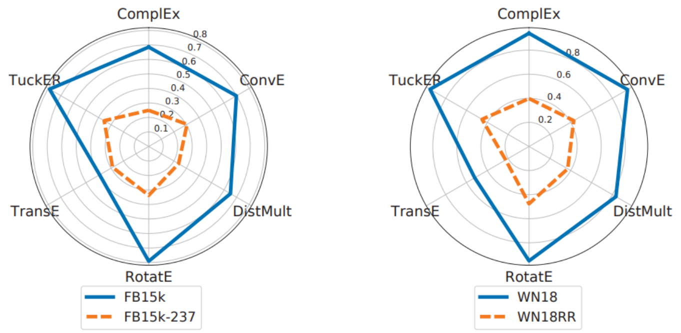


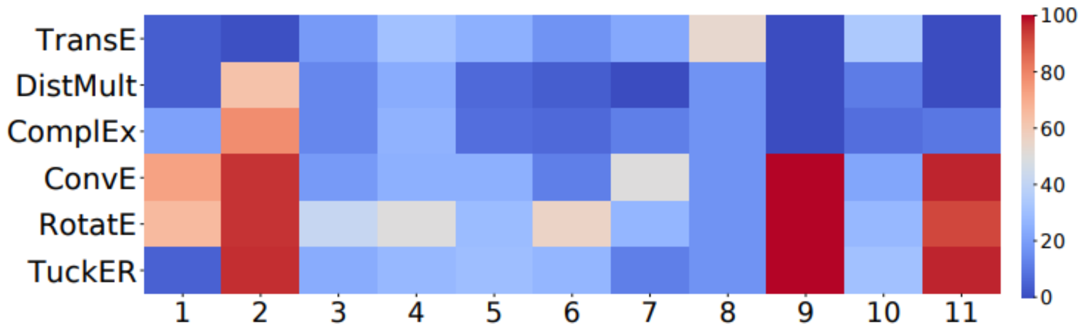


This dominance effect corroborates the seminal work of Colavita (1974) in Psychophysics where it has been demonstrated that visual stimuli dominate over the auditory stimuli when humans are asked to perform a simple audiovisual discrimination task.

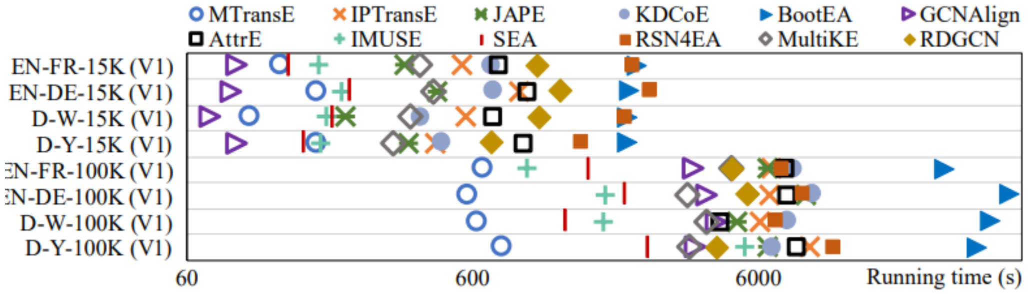
Is the multi-modal KG really helpful?

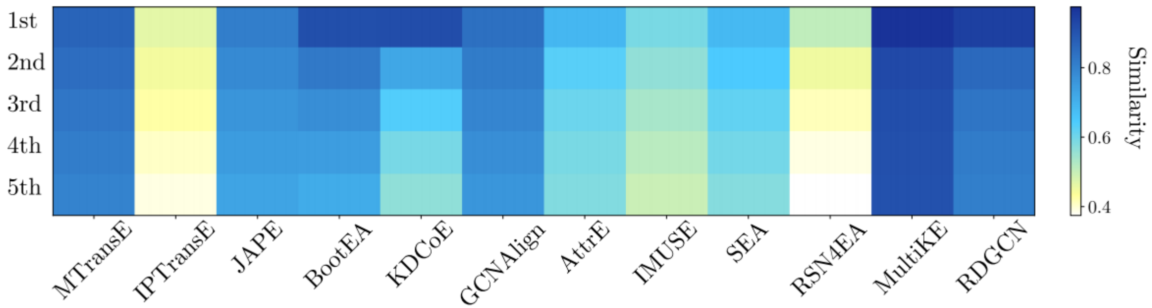
Realistic Re-evaluation of Knowledge Graph Completion Methods: An Experimental Study. SIGMOD 2020. (CCF A)



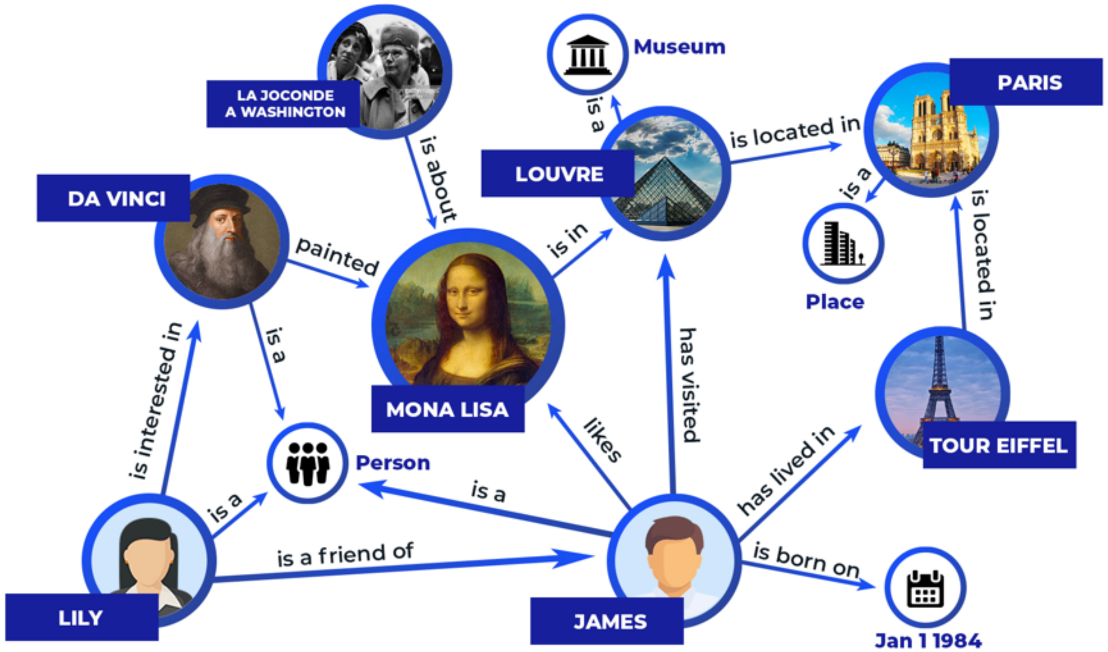


A Benchmarking Study of Embedding-based Entity Alignment for Knowledge Graphs. PVLDB 2020. (CCF A)





Is the multi-modal KG really helpful?



Cross-modal Entity Linking ? ? ?

Is the multi-modal information only needed in very specific cases for KG?

- Multimodality
- Multimodal KG Construction
- Inference
- Challenges

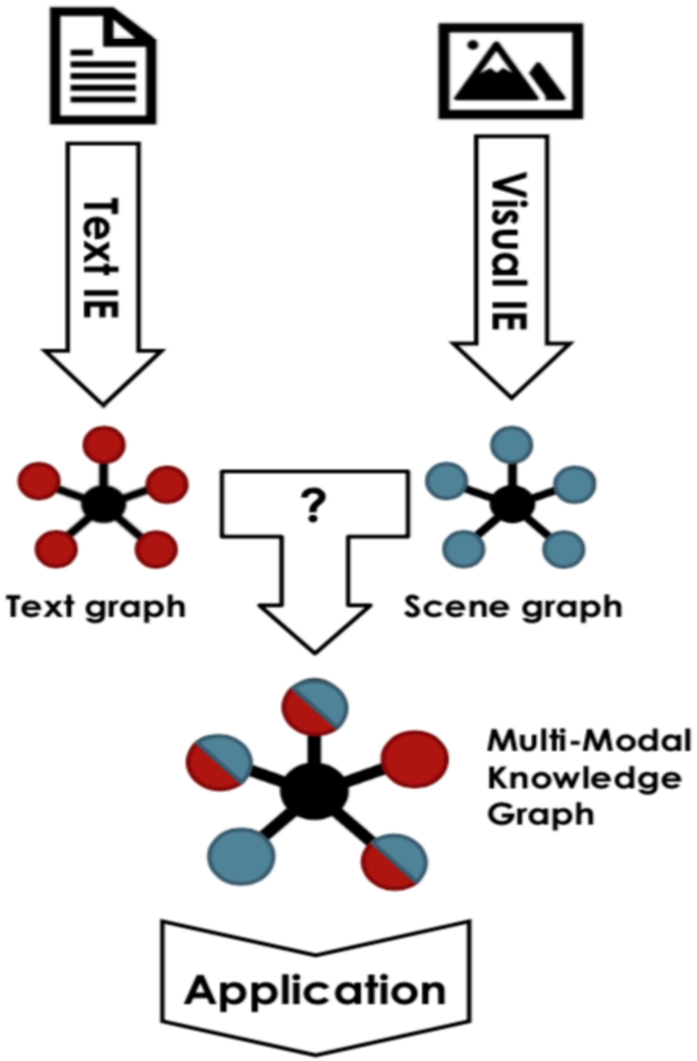
Challenges

Parsing text to structured semantic graph

Parsing images/videos to structures

Grounding event/entities across modalities

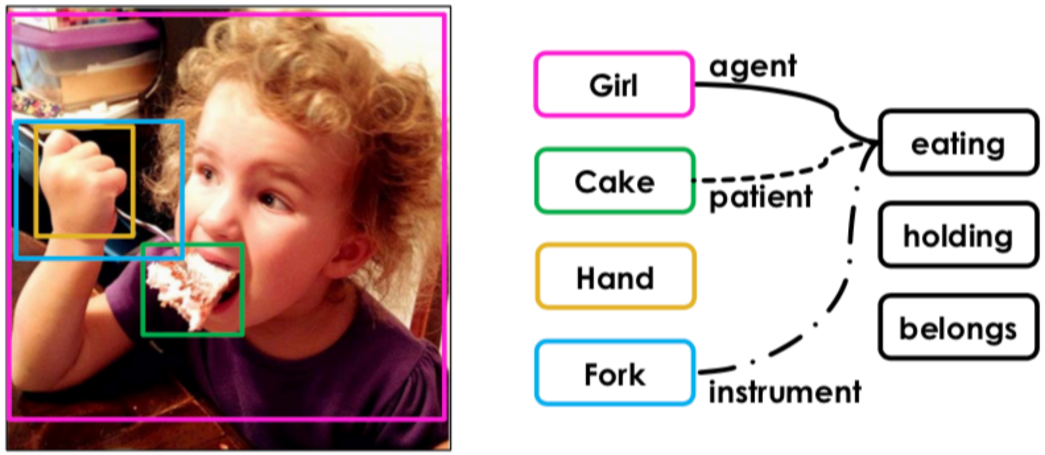
Annotation Cost or Limited training data (domain specific)
Computational complexity
Limited fixed vocabulary
Abstract concept not groundable

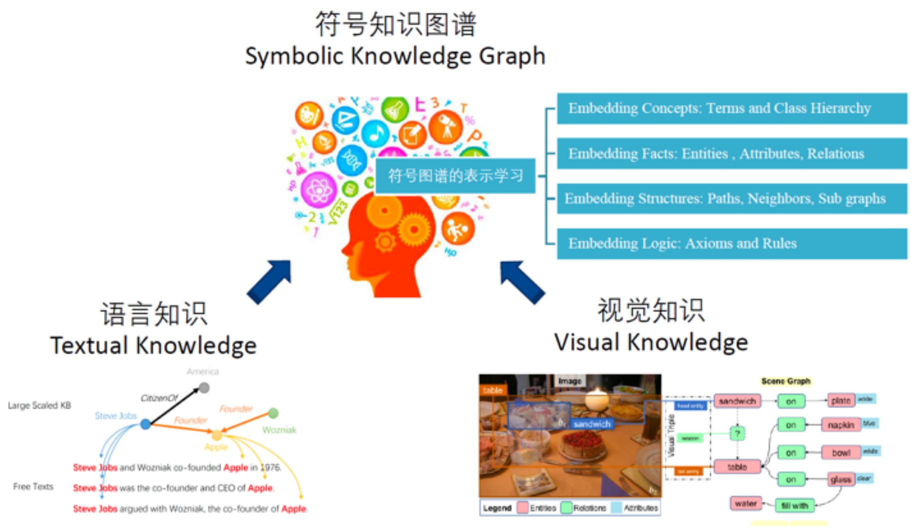


Shih-Fu Chang Alireza Zareian, Hassan Akbari, Brian Chen, Columbia University
Heng Ji, Spencer Whitehead, Manling Li, UIUC

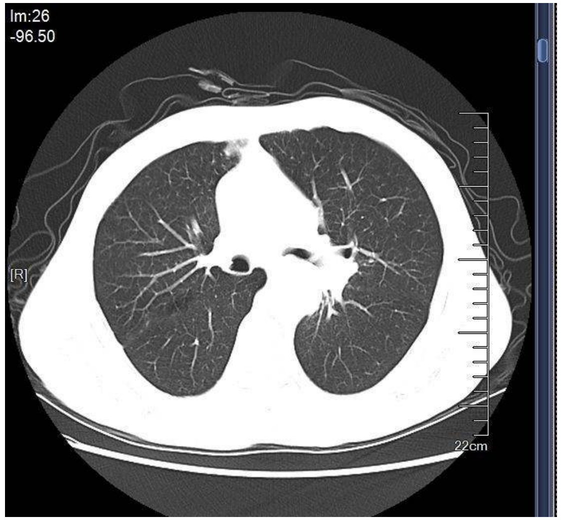
Real Challenges

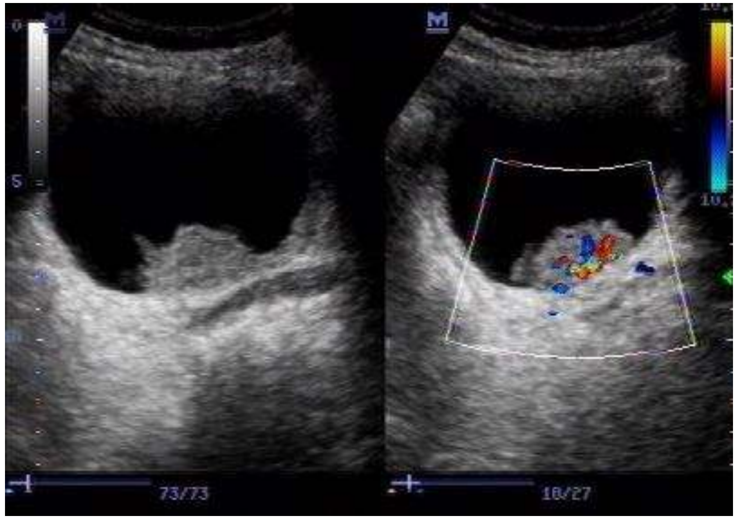
- Multimodal Data:
  - KG
  - Text
  - Image or video
- Multimodal Knowledge Representation:
  - Multimodal
  - Spatial-Temporal
  - Event
  - Rules
- Multimodal Representation Learning:
  - Pre-trained model for multimodal KG
  - Cross-modal alignment
  - Computing and storage capacity

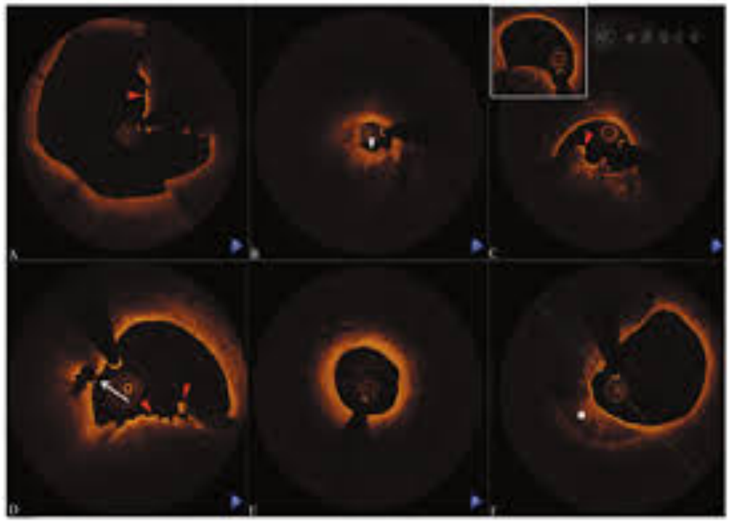


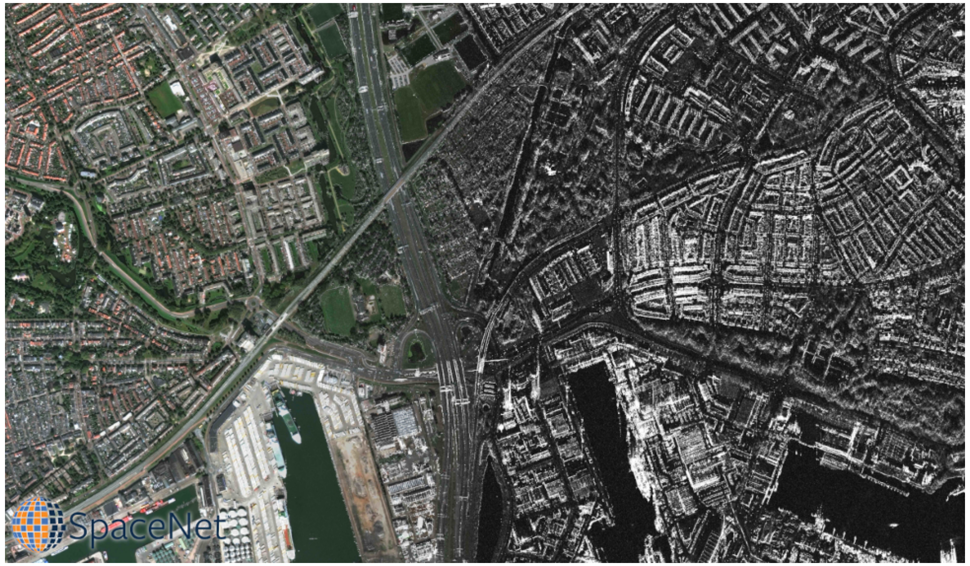


Multimodal Image?

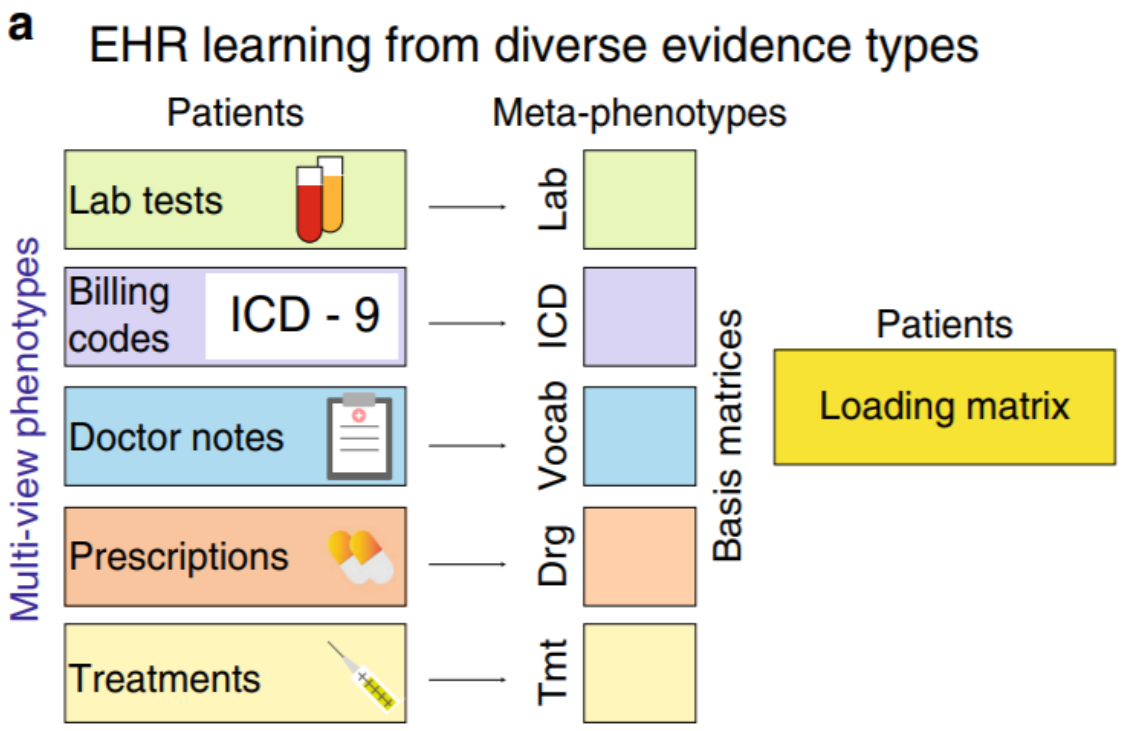


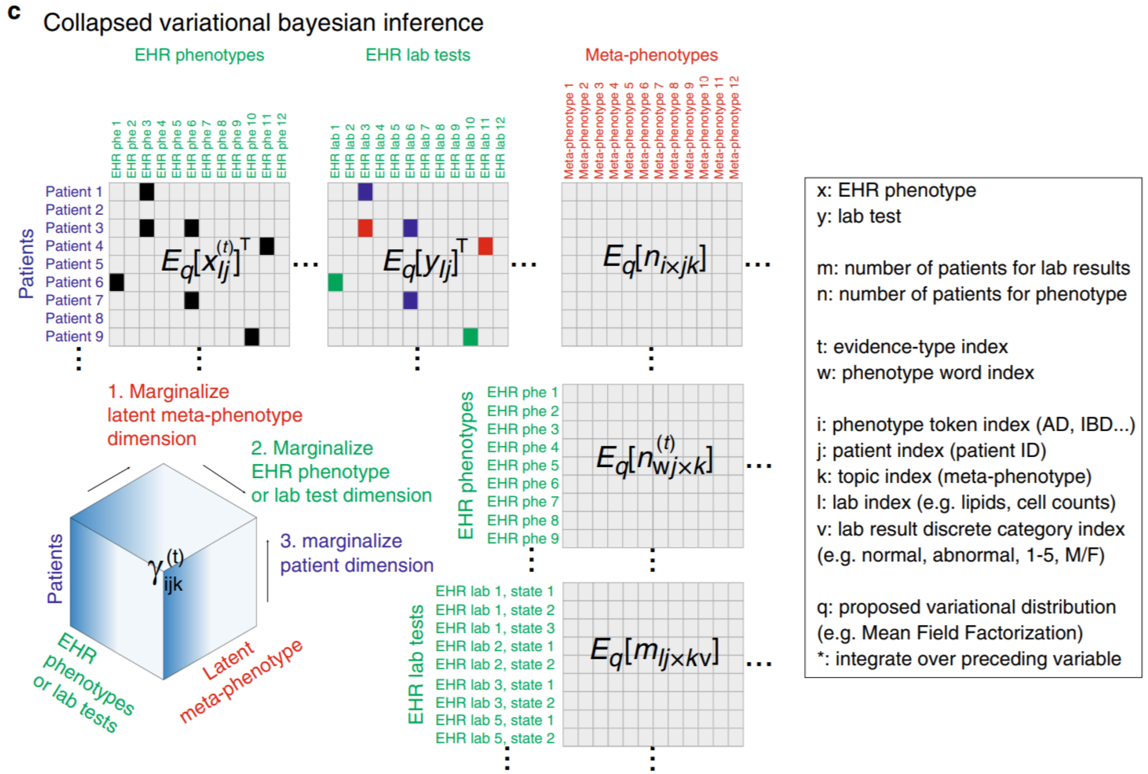






Other Multimodal Data?

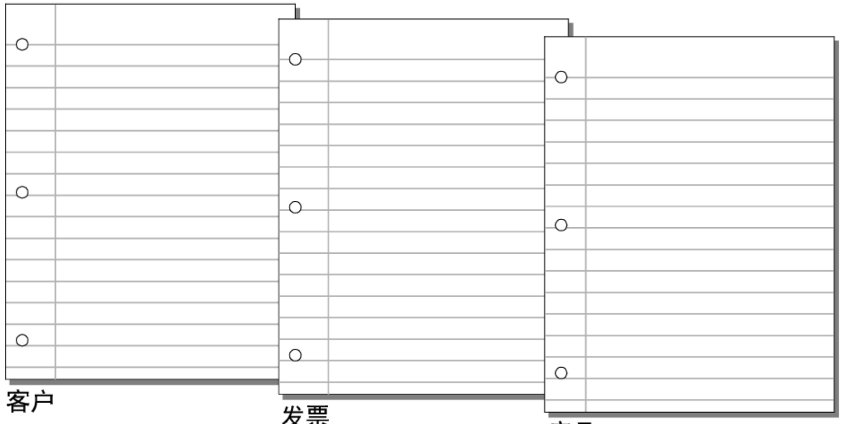




Li, Yue, et al. "Inferring multimodal latent topics from electronic health records." *Nature communications* 11.1 (2020): 1-17.

Other Multimodal Data?

Multimodal Tabular Data?



销售部	第一季度	第二季度	第三季度	第四季度	合计	辅助列 第一季度	辅助列 产品
销售1部	699	630	892	774	2995	699	销售1部
销售2部	800	988	814	796	3398	800	销售2部
销售3部	870	590	646	618	2724	870	销售3部
销售4部	1098	488	992	650	3228	1098	销售4部
销售5部	1234	458	936	690	3318	1234	销售5部
合计	4701	3154	4280	3528	15663	10962	三季度合计

Feature	Elements
Visual	Composition, Texture, Size, Color, Saturation, Focus
Motion	Zooming/Tracking, Camera Position, PerspectiveSpeed, Pan/Tilt, Editing
Audio	Volume, Speed, Pitch, Music, Tone, Frequency
Text	Size, Placement, Color, Diction, Tone, Font

Richpedia Demo

Thank you!