UKnow: A Unified Knowledge Protocol with Multimodal Knowledge Graph Datasets for Reasoning and Vision-Language Pre-Training

Biao Gong[1], **Shuai Tan**[2], **Yutong Feng**[1], **Xiaoying Xie**[1],
Yuyuan Li[3,4†], **Chaochao Chen**[4], **Kecheng Zheng**[2], **Yujun Shen**[2], **Deli Zhao**[1],
[1]Alibaba Group, [2]Ant Group, [3]Hangzhou Dianzi University, [4]Zhejiang University

{a.biao.gong, tanshuai2001, fengyutong.fyt}@gmail.com souyu.xxy@alibaba-inc.com
y2li@hdu.edu.cn zjuccc@zju.edu.cn {zkechengzk, shenyujun0302, zhaodeli}@gmail.com

Abstract

This work presents a unified knowledge protocol, called ***UKnow***, which facilitates knowledge-based studies from the perspective of data. Particularly focusing on visual and linguistic modalities, we categorize data knowledge into five unit types, namely, in-image, in-text, cross-image, cross-text, and image-text, and set up an efficient pipeline to help construct the multimodal knowledge graph from any data collection. Thanks to the logical information naturally contained in knowledge graph, organizing datasets under *UKnow* format opens up more possibilities of data usage compared to the commonly used image-text pairs. Following *UKnow* protocol, we collect, from public international news, a large-scale multimodal knowledge graph dataset that consists of 1,388,568 nodes (with 571,791 vision-related ones) and 3,673,817 triplets. The dataset is also annotated with rich event tags, including 11 coarse labels and 9,185 fine labels. Experiments on 4 benchmarks demonstrate the potential of *UKnow* in supporting common-sense reasoning and boosting vision-language pre-training with a single dataset, benefiting from its unified form of knowledge organization. See Appendix A to download the dataset.

1 Introduction

Recent efforts have been attracted to leverage the *multimodal knowledge graph* [95] for data-driven intelligence. Inspired by the human mastery knowledge network [49], we consider that the multimodal knowledge graph, which naturally accommodates heterogeneous data based on its format of complex network [93, 77], is well suited for constructing a unified knowledge criterion from the perspective of data. Driven by the multimodal knowledge graph, models can easily introduce external knowledge [57], discover long-range relations [82] and understand more logical semantics [52]. However, existing datasets of the multimodal knowledge graph commonly focus on only one task like common-sense reasoning [81, 46] due to their limited scale and irregular data organization. Therefore, it is imperative to construct a well-organized multimodal knowledge graph dataset with large-scale and rich-logic, which enables delving into deeper foundational problems in lower layers, such as the knowledge based vision-language pre-training.

To this end, we propose ***UKnow***, a **U**nified **Know**ledge protocol, which facilitates knowledge-based studies from data perspective. Particularly focusing on visual and linguistic modalities, we categorize data knowledge into five unit types, namely, in-image I_{in}, in-text T_{in}, cross-image I_{cross}, cross-text T_{cross}, and image-text IT_{cross}. As shown in Fig. 1, these knowledge types are together named as *Knowledge-View* which can be easily used to construct a multimodal knowledge graph ($\mathbf{G}_m$).

† Corresponding Author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

https://doi.org/10.52202/079017-0306

To verify that *UKnow* can serve as a standard protocol, we further set up an efficient data processing pipeline, consisting of *Phase-1/2/3*, to reorganize existing datasets into *UKnow*'s format. Please note that, this pipeline is also able to automatically extend an existing image-text dataset like LAION-5B [59] with more useful information to build a new dataset. A brief description of each *Phase* is as follows:

Phase-1: *Content Extraction*. We use pre-trained models to preprocess data and extract useful content. Note that pre-trained models can be replaced / added / disabled freely as needed. ***Phase-2:*** *Information Symbolization*. Since the results obtained in *Phase-1* (*e.g.*, images and texts) cannot be used directly for graph construction, we adopt information symbolization strategy to arrange all of them into the index in this phase. This information symbolization strategy numbers all original or generated data by a certain rule, which links the nodes from *Phase-1* to make a multimodal graph. ***Phase-3:*** *Knowledge Construction*. Two kinds of internal knowledge (I_{in}, T_{in}) and three kinds of associative knowledge ($I_{cross}, T_{cross}, IT_{cross}$) are aggregated into one graph ($\mathbf{G}_m$) in this phase as shown in Fig. 1.

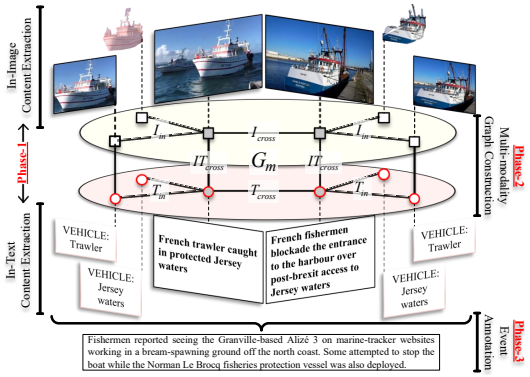


Figure 1: **Overview of *UKnow* protocol**, consisting of five unit knowledge types, namely, in-image I_{in} (*e.g.*, object), in-text T_{in} (*e.g.*, entity), cross-image I_{cross} (*e.g.*, image similarity), cross-text T_{cross} (*e.g.*, text continuity), and image-text IT_{cross} (*e.g.*, description).

Following *UKnow* protocol and above pipeline, we build a novel large-scale multimodal knowledge graph. Considering that a large-scale event dataset is of practical significance for real-world applications, such as information retrieval and public sentiment analysis, our data are collected from public international news. Overall, our dataset contains 1,388,568 nodes of which 571,791 are vision relevant (*i.e.*, news images or visual objects). The number of triples in the entire graph is 3,673,817. To the best of our knowledge, this dataset has become the largest multimodal knowledge graph dataset of international news events. Moreover, to organize data in a more structured way and enhance dataset with more category labels, our dataset introduces a *hierarchical event annotation* for each news, including ***Event-11*** and ***Event-9185***. Specifically, the former contains general event categories such as *"Sports, Ceremony, ..."*, while the latter consists of real human activity in the history such as *"2019 NBA All-Star Game, 2019 Daytona 500, ..."*. More details about the annotation are shown in Sec. 3.2, Fig 3, and Tab. 3.

In summary, our **contributions** are as follows:

- We propose *UKnow* to introduce the multimodal knowledge graph into the vision field as a new standard of data organization, which features the relation inside data in addition to the original data format. Such a protocol opens up the possibilities of data usage such that more logic-rich downstream tasks can be expected in the future.
- We design an efficient data processing pipeline for constructing dataset following our *UKnow* protocol, together with a large-scale multimodal knowledge graph dataset collected from public international news. We also equip the dataset with hierarchical event annotations, which can help models understand human activities and history. See Appendix A to download the dataset.
- We provide some examples of the usage of *UKnow* in practical applications. Experiments on four benchmarks showcase the advantages of *UKnow* in supporting common-sense reasoning and boosting vision-language pre-training with a unified form of data organization, making it possible to evaluate various tasks on a single dataset.

2 Related Work

2.1 Existing Knowledge Representation Formats

In recent years, a growing abundance multi-modal data are disseminated, linking diverse information across various modalities such as text and image in a global data space. This interconnected web of heterogeneous data constitutes a vast repository of information termed as knowledge. With the development of large-scale models, the utilization of knowledge has seen a notable surge in

https://doi.org/10.52202/079017-0306

Table 1: **Statistics** of various multimodal knowledge graph datasets. **TRIPLE** is the basic component of knowledge graph (Sec. 2.1), **WEB** and **GIT** indicate homepage and Github repository respectively. **EVENT** indicates the news event.

DATASET	YEAR	MULTIMODAL INFO.	SOURCE	NODE	IMAGE	TRIPLE	WEB	GIT	EVENT
WN9-IMG-TXT [81]	2016	ENT.	WN18, ImageNet	6,555	63,225	14,397		✓	✗
ImageGraph [47]	2017	ENT./CONCEPT	FB15k	14,870	829,931	564,010		✓	✗
VisualGenome [26]	2017	ENT.	MSCOCO	75,729	108,077	1,531,448	✓		✗
GAIA [92]	2018	ENT./CONCEPT	Freebase, Geonames	457,000	-	38,000		✓	✗
MMKG-FB15k [38]	2019	ENT./CONCEPT	FB15k, Search Engine	14,951	13,444	592,213	✓	✓	✗
MMKG-DB15k [38]	2019	ENT./CONCEPT	DB15k, Search Engine	14,777	12,842	99,028	✓	✓	✗
MMKG-YAGO15k [38]	2019	ENT./CONCEPT	YAGO15k, Search Engine	15,283	11,194	122,886	✓	✓	✗
Richpedia [72]	2020	ENT./REL./CONCEPT	Wikipedia	29,985	2,914,770	2,708,511	✓	✓	✗
VisualSem [1]	2020	ENT./CONCEPT	BabelNet	89,896	930,000	1,500,000		✓	✗
RESIN [79]	2021	ENT./REL./CONCEPT	News	51,422	6,399	150,220	✓	✓	✓
MKG-W [83]	2022	ENT./REL./CONCEPT	Open EA [67], Search Engine	15,000	14,463	-			✗
MKG-Y [83]	2022	ENT./REL./CONCEPT	Open EA, Search Engine	15,000	14,244	-			✗
MMKB-DB15K [83]	2022	ENT./REL./CONCEPT	Open EA, Search Engine	12,842	12,818	-			✗
MarKG [91]	2023	ENT./CONCEPT	Wikidata, Search Engine	11,292	76,424	34,420		✓	✗
Multi-OpenEA [36]	2023	ENT./CONCEPT	Open EA, Search Engine	920,000	2,705,688	-		✓	✗
UMVM [6]	2023	ENT./CONCEPT	DBpedia, Multi-OpenEA	238,208	1,073,671	982,626			✗
AspectMMKG [90]	2023	ENT./CONCEPT	Wikipedia, Search Engine	2,380	645,456	-		✓	✗
TIVA-KG [74]	2023	ENT./REL./CONCEPT	Wikipedia, Search Engine	443,580	1,695,688	1,382,358	✓		✗
VTKG-C [29]	2023	ENT./CONCEPT	ConceptNet, WordNet	43,267	461,007	111,491		✓	✗
UKnow	2024	ENT./REL./CONCEPT	News, Wikipedia	**1,388,568**	**1,073,671**	**3,673,817**	✓	✓	✓

exploration. Existing knowledge-based deep learning models are broadly divided into two aspects: (1) external knowledge introduction [12], (2) internal knowledge mining [22]. The former leverages expert knowledge by introducing external data [44, 28, 4] or pre-trained models [76, 58, 11, 86]. The latter means constructing correlations of training data by similarity [48, 13, 17] or discovering favorable substructures of internal models [32, 7, 33, 78].

However, from the perspective of data organization, existing studies often claim to be knowledge-based only using one piece of them, which is actually incomplete and cannot be analogous to the complex knowledge network held by humans. In this work, we build a unified knowledge protocol based on the multimodal knowledge graph to define the unified knowledge on multimodal data.

2.2 Multimodel Knowledge Graph Datasets

The Multimodal Knowledge Graph (MMKG) serves as a potent means to store and leverage multimodal knowledge explicitly, which bolsters and enhances model performances across diverse domains. In Tab. 1, we list mainstream multimodal knowledge graph datasets [72, 47, 26, 1, 92, 81, 38, 79, 83, 36, 6, 90, 74, 88, 29], constructed by texts and images with detailed information. In terms of data scale, VisualGenome [26] is a multimodal knowledge graph which contains 40,480 relations, 108,077 image nodes with objects. The ImageGraph [47] further pushed up the number of image nodes to 829,931 but missing the extraction of visual objects. Recently, VisualSem [1] implements a multimodal knowledge graph with $938K$ image nodes and 89,896 entity nodes, but it only uses 15 types of relation to build the graph. On the route of increasing the number of entity nodes, while Multi-OpenEA [36] boasts 920,000 entity nodes, surpassing prior methods, our endeavor has achieved 1,388,568 nodes, establishing the largest graph thus far. Besides, most of existing multimodal knowledge graphs are more like a vision-similarity-based image library [40, 65] with image descriptions and meta information, it lacks the most valuable feature of the knowledge graph: "The Logical Connection". This logic refers to the additional association between two nodes that were originally unrelated, triggered by a news event involving these two nodes. For example, prior to the news event "Celebrity 1 visits Area 1," there was no relation between Celebrity 1 and the Area 1. The newly added "visit" relation in <("`Celebrity1`"), `visit`, ("`Area1`")> tuple exemplifies this logic, which is highly beneficial for downstream tasks.

Generally speaking, the above news refer to international news, which carries the most complex event logic as well as plentiful multimodal information [75]. To completely exploit the advantages of multimodal knowledge graphs, building a dataset using event logic from international news is a natural approach. However, there is not yet a large multimodal knowledge graph of news events. RESIN [79] is a recently published multimodal knowledge graph containing 24 types of entities, 46 types of relations and 67 types of events. The larger and fresher CLIP-Event [33] is a event rich dataset with 106,875 images and 187 types of events extracted by a text information extraction system [92, 37]. Actually, CLIP-Event is not a knowledge graph and its definition of "event" is not a news event but an action. In summary, one of goals of our work is to build a large, and realistic news-event rich, multimodal knowledge graph dataset from international news.

https://doi.org/10.52202/079017-0306

2.3 Knowledge-based Downstream Tasks

Thanks to the innovative unified knowledge proposed by our *UKnow* protocol, our dataset can readily accommodate a variety of downstream tasks. In this study, we opt for common-sense reasoning and vision-language pre-training as experimental domains to validate our dataset. Common-sense reasoning is an extremely popular task in the field of knowledge graph. Since our dataset is based on the knowledge graph, the performance validation on common-sense reasoning is indispensable. Moreover, the representations from Vision-Language Pre-training models are capable of diminishing the necessity for intricate task-specific architectures [9], which allows the knowledge to further flow into various downstream tasks. By incorporating these two tasks, we are able to maximize the assessment of the dataset's knowledge validity.

Common-sense Reasoning. Common-sense reasoning means answering queries by logic permutations. The specific task in this work is the link prediction. Various works [3, 70, 68, 60, 94, 53] achieve reasoning by embedding entities and relations in knowledge graph into low-dimensional vector space. Path-based methods [27, 82, 63, 51] start from anchor entities and determine the answer set by traversing the intermediate entities via relational path. There are also GCN [25] based methods [61, 16] pass message to iterate graph representation for reasoning.

Vision-Language Pre-training Vision-language pre-training (VLP) can be divided into three categories based on how they encode images [10]: OD-based region features [5, 31, 34, 41, 66, 69], CNN-based grid feature [62, 19, 20] and ViT-based patch features [84, 30, 24]. Pre-training objectives are usually: masked language/image modeling (MLM/MIM) [2, 9, 39], image-text matching (ITM) [34, 19, 10], and image-text contrastive learning (ITC) [30, 50, 35].

3 UKnow

We commence by introducing the overall architecture of *UKnow* in Sec. 3.1. Then the detailed exposition of the data collection process for the new dataset and statistics are presented in Sec. 3.2 and Sec. 3.3. In Sec. 4, we lastly provide the guidance to researchers on how to integrate the multimodal knowledge graph and effectively design a UKnow-based model.

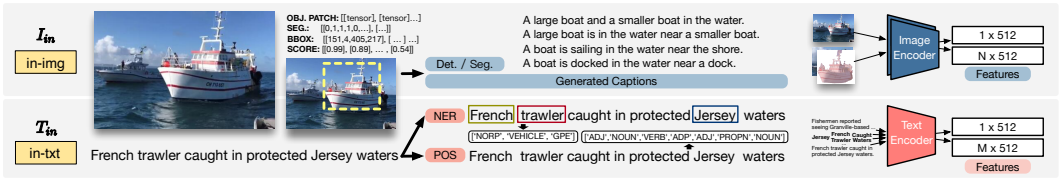
Compared to previous libraries-like methods [40, 65] with simple descriptions and meta-information, which lack the logical connection, the most valuable feature of our data processing pipeline is to endow with more logical connections to achieve superior performance in various tasks. As shown in Fig. 2, particularly focusing on visual and linguistic modalities, we categorize data knowledge into five unit types. Then we devise an efficient data processing pipeline to help reorganize existing datasets or create a new one under *UKnow* format. The construction process of *UKnow* can be invoked separately for any multimodal data to standardize the knowledge. As shown in Fig. 3, the whole pipeline is mainly empowered by three parts: *content extraction*, *information symbolization*, and *knowledge construction.*

3.1 Construction Pipeline for UKnow Protocol

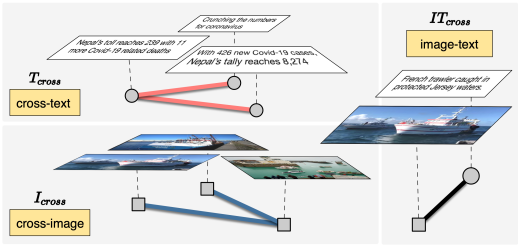
Phase-1: Content Extraction. *Content Extraction* is used to extract useful information from different fields by pre-trained deep learning models. The pre-processing functions are designed as $\mathbf{P} = \{P_1, P_2, \ldots, P_k\}$. Note that $\mathbf{P}$ can be replaced / added / disabled freely as needed. We choose pre-trained models with both global descriptions and semantic level granularity:

$$\mathbf{P} = \begin{cases} \{P_1, P_2\}, & Image\ Encoder\ [50, 18] \\ \{P_3, P_4, P_5\}, & Image\ Caption\ [42, 45, 8] \\ \{P_6\}, & Image\ Det./Seg.\ [80] \\ \{P_7\}, & Text\ Encoder\ [50] \\ \{P_8\}, & Text\ NER/POS\ [73] \\ \{P_9\}, & Annotation \end{cases} \tag{1}$$

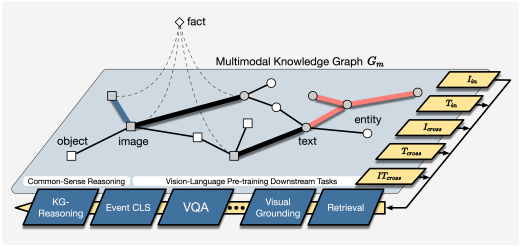
where *Det. / Seg.* and *NER / POS* refer to *Detection / Segmentation* and *Named Entity Recognition / Part-of-Speech tagging*. Then we construct the $N_p^{ori} = \mathbf{P}(I, T)$ (I is a RGB-image and T is a text) which contains a wealth of external knowledge. At this stage, all inputs concurrently go through the entire $\mathbf{P}$. It also supports the combined use of pre-trained models such as $P_6 \rightarrow P_2$ (*e.g.*, extracting the features of each object detected from the image). The final output of *Content Extraction* can be formulated as $N_p = Merge(N_p^{ori})$. $Merge$ transforms the



(a) I_{in} and T_{in}.



(b) I_{cross}, T_{cross} and IT_{cross}.



(c) Illustration of the complete UKnow.

Figure 2: **Detailed data organization under *UKnow* protocol**, which builds the multimodal (image & text) graph $\mathbf{G}_m$ based on the *Knowledge-View* (I_{in}, T_{in}, I_{cross}, T_{cross}, and IT_{cross}). Each node owns up to 22 attributes shown as N_p in Fig. 3.

original output N_p^{ori} into a K:V dictionary N_p. The KEY of N_p are shown in top right corner of Fig. 3 (N_p [*Phase-1*]). N_p is also used as the attribute of each node in the final output multimodal knowledge graph $\mathbf{G}_m$.

Phase-2: Information Symbolization. Since Images and texts cannot be used directly for graph construction, we design the *Phase-2* to number all original or generated data by a certain rule, then *Phase-3* links these nodes to make a multimodal graph. *Information Symbolization* is used to subscript N_p to edge index $\mathbf{N}_e$ or node index $\mathbf{N}_n$: (1) The symbolization for edges $\mathbf{N}_e$ is based on the category or visual / semantic similarity. For example, "[111] title_title_clip" is a kind of parallelism edge which is constructed by the cosine similarity of clip features of news titles. (2) The symbolization for nodes $\mathbf{N}_n$ is divided into three levels: [fact, image / text, object / entity]. As shown in Fig. 3, $[L_1.*]$ means fact-level which is an abstraction of a piece of news. The real index used in our multimodal knowledge graph would be $\{L_1.0, L_1.1, L_1.2, ...\}$. Similarly, $[L_2.*]$ means image / text-level which is the symbolization of images or texts from news, $[L_3.*]$ is the object in image or entity in text. The index for all nodes is eventually shuffled, that is, the real index would be $\{L_1.0, L_2.1, L_1.2, L_3.3, L_3.4, ...\}$.

We provide the clearer explanations about the motivation of Phase-2. As stated in Sec. 3.2, our data are collected from international news, which encompasses a wide variety of text and images. Although Phase-1 preprocesses the data like detection and segmentation, the resulting features are still a huge volume as it contains detailed information extracted from the news. While this detailed information is valuable for constructing a knowledge graph, the computational demands and complexity far exceed available resources. Thus, a common approach in knowledge graph construction is to store data and their relationships as indices, as done in the Phase-2 Information Symbolization stage. This means it has the following benefits: efficiency in storage and retrieval, fast lookup and traversal, uniqueness and consistency, scalability, and simplification of graph operations.

Phase-3: Knowledge Construction. We categorize data knowledge into five unit types, namely, in-text (T_{in}), in-image (I_{in}), inter-text (T_{cross}), inter-image (I_{cross}), and image-text (IT_{cross}) which are together called *Knowledge-View* detailed in Fig. 2(a) and Fig. 2(b).

Table 2: **Edge ($\mathbf{N}_e$) construction and statistics.**

Phrase	Construction Method	View	Num.
Phrase-2	Detection Category	I_{in}	648,871
	NER Category	T_{in}	1,606,936
	Similarity&Manual Annotation	IT_{cross}	684,207
	Similarity&Manual Annotation	T_{cross},I_{cross}	140,133
Phrase-3	Manual Event Annotation	-	593,670

In this phase, we aggregate two kinds of internal knowledge (I_{in}, T_{in}) and three kinds of associative knowledge ($I_{cross}, T_{cross}, IT_{cross}$) in one graph $\mathbf{G}_m$, which are usually introduced independently in previous studies. *Knowledge Construction* takes as input the edge index $\mathbf{N}_e$ and node index $\mathbf{N}_n$ numbered by *Phase-2* and output the multimodal

https://doi.org/10.52202/079017-0306

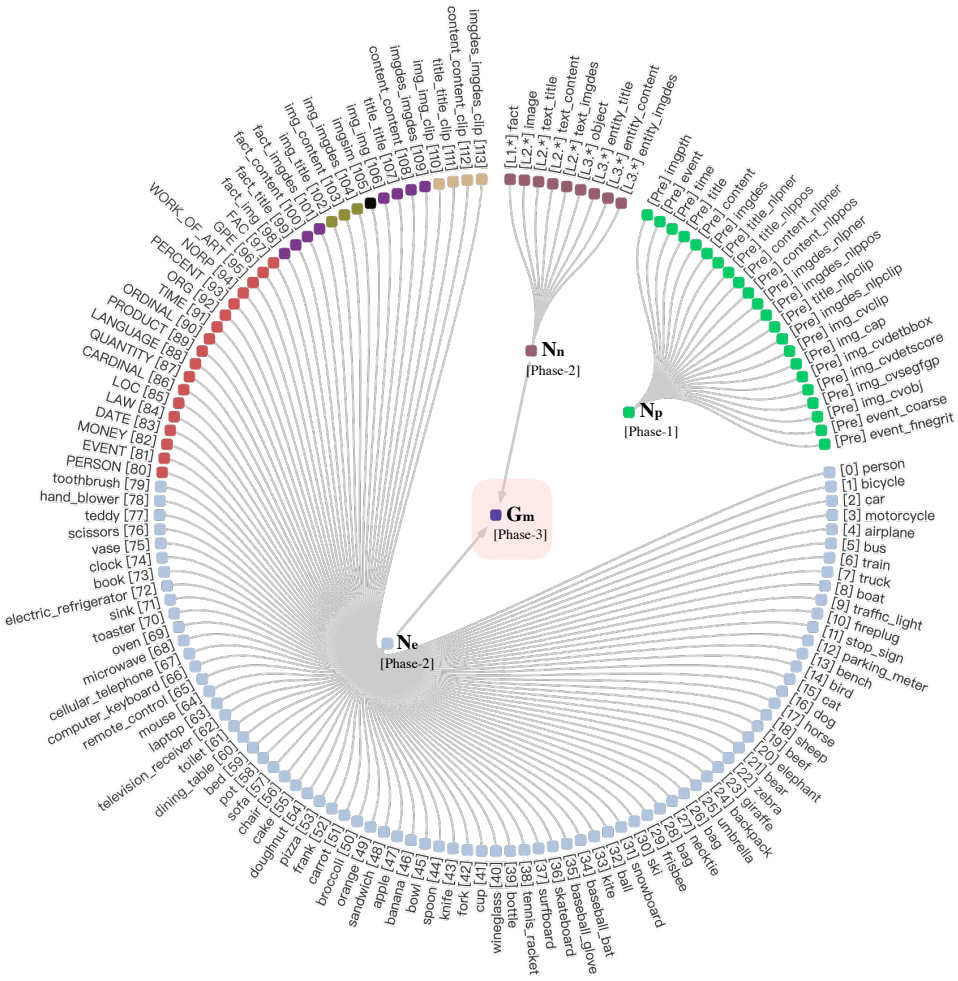


Figure 3: **Pipeline of dataset construction following *UKnow* protocol.** Phase-1: *Content Extraction* (N_p), Phase-2: *Information Symbolization* ($\mathbf{N}_n$, $\mathbf{N}_e$), and Phase-3: *Knowledge Construction* ($\mathbf{G}_m$). $\mathbf{N}_n$ hides the real node index for easy understanding, the actual number is much more than $\mathbf{N}_e$.

knowledge graph $\mathbf{G}_m$ (Fig. 2(c)). Since $\mathbf{N}_e$ and $\mathbf{N}_n$ are both isolated, we use four kinds of correlation methods including semantic similarity, visual similarity, annotations, and categories to make connections between $\mathbf{N}_n$ by $\mathbf{N}_e$ shown in Tab. 2.

3.2 Dataset Collection

Following the proposed protocol and three phases, we collect a new dataset, a large-scale multimodal knowledge graph from public international news. Specifically, based on the Wikipedia API [43] and our crawler system, we grab all the data of "*Worldwide Current Events*" from Wikipedia. As demonstrated in top of Fig. 4, we propose two category sets of news event called: *Event-11* and *Event-9185*, which is coarse-grained and fine-grained respectively. For example, *"Sports"* is a kind of coarse-grained event label in *Event-11* and *"2019 Daytona 500"* is a fine-grained label in *Event-9185*, detailed in Tab. 3. Since Wikipedia only records the news URL (downward black arrow in Fig. 4) and the HTML of original news from different news platforms is inconsistent, it is difficult to design a uniform crawler to get the well-structured raw data of news. Thus, we manually read each news and collect the original data (rightward black arrow). By this way, each news in our dataset is marked with extremely clean **title**, **content**, **time**, [**image**], **image description**, **event description**, [hierarchical] **event name** (*e.g.*, *"Armed conflicts and attacks→War in Donbass"*), and **event attribute** (location, date, *etc*). Subsequently, as shown in bottom right of Fig. 4, we apply the designed pipeline to sequentially undergo phases 1/2/3 to restructure the above extracted raw data, resulting in the knowledge graph under the *UKnow* format.

https://doi.org/10.52202/079017-0306

Figure 4: **Event category** labeled on web data and the **data flow diagram.**

Table 3: **Details of the event category.**

Event Name (*Event-11*)	Visual	Texual	All	Event Name (10 examples of *Event-9185*)	Visual	Texual	All
Armed conflicts and attacks	87,346	90,157	177,503	Saudi Arabian-led intervention in Yemen	555	258	813
Arts and culture	11,059	14,896	25,955	A boat carrying Indonesian migrants capsizes off the southern coast of Malaysia	46	19	65
Business and economy	12,598	25,565	38,163	Travel restrictions related to the COVID-19 pandemic	753	796	1,549
Disasters and accidents	28,062	47,459	75,521	GameStop short squeeze	45	175	220
Health and environment	230,926	258,349	489,275	Opposition to Brexit in the United Kingdom	383	93	476
International relations	37,349	56,444	93,793	Gretchen Whitmer kidnapping plot	167	308	475
Sports	15,647	31,194	46,841	Legality of euthanasia	185	455	640
Law and crime	69,573	86,514	156,087	Ukraine International Airlines Flight 752 (Air Crash)	314	179	493
Politics and elections	74,477	72,714	147,191	Manhattan blackout	269	90	359
Science and technology	4,062	15,556	19,618	2019 Lagos school collapse	524	119	643
Others	236	184	420	...	...	...	...

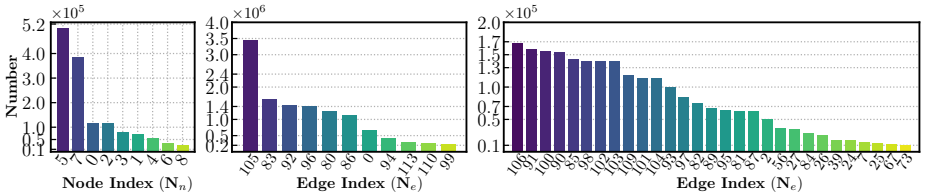
Table 4: **Partition of our dataset.**

PARTITION	T_{in}		I_{in}		T_{cross}		I_{cross}		IT_{cross}	
	NODE	EDGE	NODE	EDGE	NODE	EDGE	NODE	EDGE	NODE	EDGE
Training Set	448,691	8,030,531	501,564	979,287	250,858	396,200	69,911	421,628	765,654	382,827
Validation Set	37,488	100,280	12,126	12,212	69,533	57,162	15,532	97,272	9,764	4,882
Testing Set	37,668	100,375	12,182	12,261	69,286	55,464	15,336	99,303	9,622	4,811
Pre-training Set	228,339	435,659	343,458	325,755	101,880	314,918	47,017	271,593	278,058	139,029
Fine-tuning Set	75,924	82,350	65,809	61,850	19,185	59,832	8,880	52,772	52,522	26,261
Testing Set	34,422	28,219	22,809	22,278	6,633	21,360	3,074	17,754	18,186	9,093

Furthermore, in addition to utilizing intricate annotation files (*e.g.*, Fig. 4) as inputs mentioned above, another major advantage of the proposed conversion pipeline is its ability to accommodate common image-text pair annotations expressed in the format of *"[image description]* \t *./xxx.jpg* \n*"*), as the fundamental input. This design allows *UKnow* to automatically construct a new dataset with more useful information from an existing image-text pair dataset. Taking LAION-5B [59] as an example, which solely comprises pairs of images and text, our pipeline can extract more features from them like objects, and thus expand LAION-5B into a larger and more practical dataset. However, given the absence of high-level event logic, this type of input does not lend itself to the creation of $[L_1.*]$ nodes and event-related edges.

https://doi.org/10.52202/079017-0306

Table 5: **Histogram of the number of indexes in our dataset**. The x-axis in the upper left corner (Node Index) corresponds to the order of the I_n in Fig. 3.



3.3 Dataset Statistics and Visualizations

Through data collection and processing in Sec. 3.2, our dataset comprises 1,388,568 nodes, of which 571,791 are relevant to vision (i.e., pertaining to a news image or a visual object), and 3,673,817 triples. The partitioning of our dataset is presented in Tab. 4, with all partitions being randomly sampled. Moreover, as depicted in Fig. 5, we present the histogram of all indices in *UKnow*. Considering our dataset is a multimodal knowledge graph, *i.e.*, each node corresponds to a multimodal data, and each edge serves the purpose of connecting either single or cross-modal nodes.

The top-2 number of nodes are "[L_3.*] objects" (501,880) and "[L_3.*] entity_content" (386,561) which belong to I_{in} and T_{in} respectively. The former represents visual objects extracted from images, and the latter means text entities extracted from news contents. The maximum number of edges is "[105] imgsim" (3,447,990) which is a kind of associative knowledge from I_{cross}.

Fig. 3.3 shows the variation in different thresholds. τ indicates the threshold of cosine similarity which controls whether edges are built between nodes. It can be adjusted according to needs (*e.g.*, storage, computational complexity, fineness). The default setting of τ is 0.8. ρ_m indicates the average edge number of connections per node in the entire graph, which demonstrates the density of a graph. As shown in Tab. 8, the whole graph is sparse with *ENTITY* as the main nodes of the background, and the subject element of the dense region is *CONTENT*. ρ means the number of edges, *i.e.*, there are 615k nodes with 0-edge or 1-edge, 463k nodes with 2-edges or 3-edges, and so on. The Mean Density (ρ_m) in Fig. 3.3 is calculated as a weighted average of ρ and the number of nodes.

Table 6: **Histogram of the number of other edge indexes.**

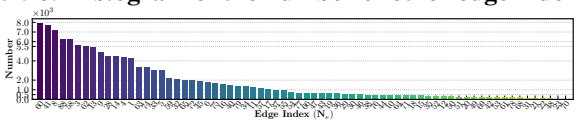


Table 7: **The variation in different similarity thresholds.**

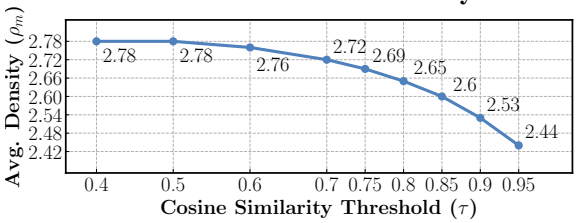


Table 8: **The graph density and mainstream node types.**

τ	**Density (ρ)**	0,1	2,3	4,5	6,7	8,9	10,11	12,13	14,15	16,17	⩾18
0.8	**NodeNum.**	615k	463k	132k	71k	55k	32k	14k	4k	703	78
	MainType	Entity	Object	Title	Image	Content					

Tab. 3 shows all the categories in *Event-11*, and 10 examples in *Event-9185*. "Visual" means the number of nodes belonging to images or objects. "All" means the number of all nodes marked with this event category. Generally speaking, *Event-9185* is specific to an exact human activity and can be used to learn the semantic relevance of news contents, while *Event-11* is more like a categorization of news events, which is benefit for archiving news materials through a trained classification model.

3.4 Why UKnow protocol

In most of existing Knowledge Graph (KG), data organization is typically driven by the requirements of specific tasks, such as image classification, object detection, or semantic search. For example, VisualGenome organizes data by creating dense image annotations linked to textual descriptions, while MMKG-YAGO15k focuses on aligning textual concepts with visual data. This task-specific data organization methods often lack a cohesive framework for organizing data across multiple modalities in a unified manner. As a result, most existing KGs are often designed with specific applications in mind, such as common-sense reasoning, multimodal event classification, or visual

https://doi.org/10.52202/079017-0306

task adaptation. While these applications benefit from the structured nature of KGs, the underlying datasets may not be flexible enough to support a wide range of tasks, particularly those that require cross-modal reasoning or dynamic context adaptation.

In contrast, UKnow is inherently designed to handle multimodal data (e.g., images, text) in a unified structure. This allows for seamless integration and interaction between different types of data, making it particularly well-suited for tasks that require cross-modal reasoning, such as vision-language pre-training and complex event understanding. Besides, UKnow introduces a hierarchical structure that organizes nodes into levels and incorporates logical connections between them. The unified structure and logical richness of UKnow make it highly versatile for a wide range of downstream tasks. The ability to evaluate various tasks on a unified Knowledge Graph also reduces the complexity of model development and evaluation, leading to more efficient and effective AI solutions.

4 Usage of UKnow

4.1 UKnow for Common-sense Reasoning

Since *UKnow* is reasoning compatible, *i.e.*, it naturally supports all KG-reasoning models, we directly implemented the commonly used KG-reasoning models (*e.g.*, TransE [3], Q2B [54]) on *UKnow*. We propose a plug-in module which aggregates node features within a small sub-graph region to achieve a better central node features. We briefly introduce how to implement this module. Suppose $N(e) \equiv \{e_{neib} | r(e_{neib}, e) \vee r(e, e_{neib}), r \in \mathcal{R}\}$ is the collection of neighbors of each central node e. The calculation expression of the new representation e' of e is as follow:

$$\mathrm{e}' = \mathrm{MLP}(Flatten(\mathrm{ReLU}(\omega_n \star (\tau'(\mathrm{e}, N_\mathrm{e}{}') + \mathrm{b}_n)), \tag{2}$$

where $\mathrm{e} \in \mathbb{R}^d$ is the node feature before enhancement, e' is the new feature, $\star$ denotes a 2D convolution operation, ω_n is the filter, b_n is the bias and the specification of MLP is $\mathbb{R}^{m_1 \times m_2} \times \mathbb{R}^d$. The concat function $\tau'(\mathrm{e}, N_\mathrm{e}) \in \mathbb{R}^{m_1 \times m_2}$ as $[\mathrm{e}; \mathrm{e}_{neib}{}^{1'}; \mathrm{e}_{neib}{}^{2'}; \ldots; \mathrm{e}_{neib}{}^{m}]$ where $\mathrm{e}_{neib}{}^{i} \in N'_\mathrm{e}$.

4.2 UKnow for Vision-Language Pre-training

Following the recent works [33], our work applies CLIP [50] as the pre-trained backbone benefit from its strong downstream performance. Specifically, the text encoder first tokenize the input text description into the word sequence, and then projects them into word embeddings $\mathbf{W}_0 = \{\mathbf{w}_0^1, \mathbf{w}_0^2, \cdots, \mathbf{w}_0^N\} \in \mathbb{R}^{N \times d^t}$. $\mathbf{W}_0$ is fed into a L-layer Transformer [71] with the architecture modifications described in BERT [9]. And the final text embedding $\mathbf{z}^T$ is obtained by projecting the last token, which corresponds to the `[EOS]` (the end of sequence) token, from the last layer of the text encoder, *i.e.*, $\mathbf{z}^T = \texttt{TextProj}(\mathbf{w}_L^N), \mathbf{z}^T \in \mathbb{R}^d$. As for the vision encoder, the input image I is first split into M non-overlapping patches, and projected into a sequence of patch tokens $\mathbf{E_0} \in \mathbb{R}^{M \times d^v}$. Then, $\mathbf{E_0}$ is fed into a L-layer Transformer-based architecture along with a learnable `[CLS]` token $\mathbf{c}_0$. The final image embedding $\mathbf{z}^I$ is obtained by projecting the `[CLS]` token from the last layer of the vision encoder, *i.e.*, $\mathbf{z}^I = \texttt{VisProj}(\mathbf{c}_L^v, \mathbf{E}_L^v)), \mathbf{z}^I \in \mathbb{R}^d$. Since we have *Knowledge-View*, a new dimension $\mathbf{z}^k$ which is used to represent knowledge is introduced:

$$\mathbf{z}^k = \texttt{Concat}(I_{in}(\mathbf{z}^I), T_{in}(\mathbf{z}^T), I_{cross}(\mathbf{z}^I), T_{cross}(\mathbf{z}^T)), \tag{3}$$

where $I_{in}(\cdot)$ and $T_{in}(\cdot)$ mean to get the embedding of the $[L_3.*]$ nodes $(\mathbf{N}_n)$ from $\mathbf{G}_m$ via $\mathbf{N}_e$, $I_{cross}(\cdot)$ and $T_{cross}(\cdot)$ mean to get the embedding of $[L_2.*]$ from $\mathbf{G}_m$. Therefore, the similarity score between the image, text and knowledge can be calculated with the cosine similarity as follow:

$$s(T, I, k) = \frac{\mathbf{z}^{T\top}\mathbf{z}^I}{\|\mathbf{z}^T\|\|\mathbf{z}^I\|} + \frac{\mathbf{z}^{k\top}\mathbf{z}^I}{\|\mathbf{z}^k\|\|\mathbf{z}^I\|} + \frac{\mathbf{z}^{k\top}\mathbf{z}^T}{\|\mathbf{z}^k\|\|\mathbf{z}^T\|}. \tag{4}$$

4.3 UKnow Baseline

Upgrading AI from understanding objects (*e.g.*, an apple) as in most current vision tasks to understanding complex human activities (*e.g.*, an event), to understanding the logic between entities or objects, and to achieving higher-order intelligence, is always the thing we would like to pioneer. Thus, in this section, we naturally present a series of novel logic-rich downstream tasks as the baselines

https://doi.org/10.52202/079017-0306

for our dataset. Specifically, Common-sense Reasoning is a conventional and fundamental task in our domain, aligning closely with our dataset. Then we perform multiple downstream tasks to verify the performance of the pretrained model trained with our dataset. For more details about task description/training setting/evaluation metric/analysis, please refer to Sec. C.

Common-sense Reasoning. We implement the Q2B* with our *UKnow* based plug-in module based on Q2B [54] and BETAE* based on BETAE [55]. As shown in Tab. 9, BETAE* achieves on average **21.64%** and **21.23%** MRR on the validation and testing set of our dataset. It indicates that our *UKnow* based module can significantly improve the performance of existing methods.

Multimodal Event Classification. As shown in Tab. 10, TCL [85] achieves on **66.80%** and **55.87%** on ACC@1 when using the image-input on the *Event-11* and *Event-9185*. respectively. We add a late-fusion module after the image/text encoder for all methods to support multimodal classification. Results show that TCL obtains gains of **1.89%** and **5.02%** compared with the singlemodal input, which demonstrates that multimodal pre-training is more helpful for downstream multimodal tasks.

Single- & Cross-Modal Retrieval. As shown in Tab. 11, TCL [85] achieves on **33.24%**, **43.37%** and **45.22%** R@1, R@5, R@10 on the zero-shot setting of image retrieval. The results are **58.89%**, **68.47%** and **73.91%** when fine-tuning the pre-trained parameters, which means the pre-training→fine-tuning strategy is extremely beneficial for downstream retrieval.

Visual Task Adaptation. As shown in Tab. 12, our approach obtains gains of avg. **1.14%** compared with the origin CLIP when fairly using the same *UKnow*'s data for the upstream pre-training. It is essential to highlight that the image-text PAIR constitutes only one type of data in our protocol. By leveraging the capabilities of *UKnow*, our pre-trained CLIP model can effectively comprehend the inherent knowledge, resulting in superior performance than original CLIP model (Tab. 12, Row2).

4.4 Practical Applications in Other Domains

UKnow is a general multimodal knowledge graph construction protocol that can be easily adapted to different domains by adjusting P in Phase-1 to the relevant processing modules required. Due to issues such as time and effort and difficulty of data acquisition, in this paper, we only use international news as an example, given its significance in the multimodal field and its ability to highlight UKnow's strengths in handling multimodal data. In the future, as we mentioned in Sec. D.3, we aim to diversify modalities by augmenting our dataset with a broader range of modalities (e.g., audio, video, 3D, etc.) to facilitate exploration across various downstream tasks. Here's an example of how to extend UKnow to the video domain and modality: (1) Phase-1: Replace P with operations like Video Captioning, Action Recognition, Video Summarization, or Object Detection and Tracking to process the video content. (2) Phase-2: Organize the processed video features into the node index and construct relationship edges with other modalities such as text, images, and audio. (3) Phase-3: Utilize Phase 3 to build the knowledge graph, which can then be applied to various knowledge-based downstream tasks.

5 Conclusion

This paper presents a unified knowledge protocol called *UKnow* to establish the standard of knowledge from the perspective of data. Following this protocol, we collect a novel and the largest multimodal knowledge graph dataset from public international news with rich news event annotations, which can help intelligent machines understand human activities and history. The specific tasks addressed in this paper are the common-sense reasoning and vision-language pre-training. The former is a typical task in the knowledge graph field, and the latter brings knowledge to various downstream tasks. We also present a series of novel logic-rich downstream tasks to showcase the advantages of *UKnow*. In future work, we will continuously expand the data of different scales based on the *UKnow* protocol.

References

[1] Houda Alberts, Teresa Huang, Yash Deshpande, Yibo Liu, Kyunghyun Cho, Clara Vania, and Iacer Calixto. Visualsem: a high-quality knowledge graph for vision and language. arXiv preprint arXiv:2008.09150, 2020.

[2] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254, 2021.

[3] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In Advances in Neural Information Processing Systems, 2013.

[4] Sanyuan Chen, Yutai Hou, Yiming Cui, Wanxiang Che, Ting Liu, and Xiangzhan Yu. Recall and learn: Fine-tuning deep pretrained language models with less forgetting. arXiv preprint arXiv:2004.12651, 2020.

[5] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In European conference on computer vision, pages 104–120, 2020.

[6] Zhuo Chen, Lingbing Guo, Yin Fang, Yichi Zhang, Jiaoyan Chen, Jeff Z Pan, Yangning Li, Huajun Chen, and Wen Zhang. Rethinking uncertainly missing and ambiguous visual modality in multi-modal entity alignment. In International Semantic Web Conference, pages 121–139. Springer, 2023.

[7] Zewen Chi, Li Dong, Bo Zheng, Shaohan Huang, Xian-Ling Mao, Heyan Huang, and Furu Wei. Improving pretrained cross-lingual language models via self-labeled word alignment. arXiv preprint arXiv:2106.06381, 2021.

[8] Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui, and Mohit Bansal. Fine-grained image captioning with clip reward. In Findings of NAACL, 2022.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[10] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. An empirical study of training end-to-end vision-and-language transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 18166–18176, 2022.

[11] Sepideh Esmaeilpour, Bing Liu, Eric Robertson, and Lei Shu. Zero-shot out-of-distribution detection based on the pretrained model clip. In Proceedings of the AAAI conference on artificial intelligence, 2022.

[12] Yuxia Geng, Jiaoyan Chen, Xiang Zhuang, Zhuo Chen, Jeff Z Pan, Juan Li, Zonggang Yuan, and Huajun Chen. Benchmarking knowledge-driven zero-shot learning. Journal of Web Semantics, page 100757, 2023.

[13] Tianyu Guo, Hong Liu, Zhan Chen, Mengyuan Liu, Tao Wang, and Runwei Ding. Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, pages 762–770, 2022.

[14] Kelvin Guu, John Miller, and Percy Liang. Traversing knowledge graphs in vector space. arXiv preprint arXiv:1506.01094, 2015.

[15] Will Hamilton, Payal Bajaj, Marinka Zitnik, Dan Jurafsky, and Jure Leskovec. Embedding logical queries on knowledge graphs. Advances in neural information processing systems, 2018.

[16] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. Advances in neural information processing systems, 2017.

[17] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. Advances in Neural Information Processing Systems, pages 5679–5690, 2020.

[18] Xiangteng He, Yulin Pan, Mingqian Tang, Yiliang Lv, and Yuxin Peng. Learn from unlabeled videos for near-duplicate video retrieval. In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 1002–1011, 2022.

[19] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12976–12985, 2021.

 https://doi.org/10.52202/079017-0306

[20] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. arXiv preprint arXiv:2004.00849, 2020.

[21] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In International Conference on Machine Learning, pages 4904–4916, 2021.

[22] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. IEEE transactions on pattern analysis and machine intelligence, pages 4037–4058, 2020.

[23] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 1780–1790, 2021.

[24] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In International Conference on Machine Learning, pages 5583–5594, 2021.

[25] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907, 2016.

[26] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision, pages 32–73, 2017.

[27] Ni Lao, Tom Mitchell, and William Cohen. Random walk inference and learning in a large scale knowledge base. In Proceedings of the 2011 conference on empirical methods in natural language processing, pages 529–539, 2011.

[28] Anne Lauscher, Olga Majewska, Leonardo FR Ribeiro, Iryna Gurevych, Nikolai Rozanov, and Goran Glavaš. Common sense or world knowledge? investigating adapter-based knowledge injection into pretrained transformers. arXiv preprint arXiv:2005.11787, 2020.

[29] Jaejun Lee, Chanyoung Chung, Hochang Lee, Sungho Jo, and Joyce Whang. Vista: Visual-textual knowledge graph representation learning. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 7314–7328, 2023.

[30] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems, pages 9694–9705, 2021.

[31] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557, 2019.

[32] Liunian Harold Li*, Pengchuan Zhang*, Haotian Zhang*, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In IEEE Conf. Comput. Vis. Pattern Recog., 2022.

[33] Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. Clip-event: Connecting text and images with event structures. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16420–16429, 2022.

[34] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In European Conference on Computer Vision, pages 121–137, 2020.

[35] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. arXiv preprint arXiv:2110.05208, 2021.

[36] Yangning Li, Jiaoyan Chen, Yinghui Li, Yuejia Xiang, Xi Chen, and Hai-Tao Zheng. Vision, deduction and alignment: An empirical study on multi-modal knowledge graph alignment. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1–5. IEEE, 2023.

[37] Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. A joint neural model for information extraction with global features. In Proceedings of the 58th annual meeting of the association for computational linguistics, pages 7999–8009, 2020.

[38] Ye Liu, Hui Li, Alberto Garcia-Duran, Mathias Niepert, Daniel Onoro-Rubio, and David S Rosenblum. Mmkg: multi-modal knowledge graphs. In European Semantic Web Conference, pages 459–474, 2019.

[39] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 2019.

[40] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1096–1104, 2016.

[41] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in neural information processing systems, 2019.

[42] Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. Discriminability objective for training descriptive captions. arXiv preprint arXiv:1803.04376, 2018.

[43] Martin Majlis. Wikipedia-api. https://pypi.org/project/Wikipedia-API/.

[44] Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14111–14121, 2021.

[45] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. arXiv preprint arXiv:2111.09734, 2021.

[46] Hatem Mousselly-Sergieh, Teresa Botschen, Iryna Gurevych, and Stefan Roth. A multimodal translation-based approach for knowledge graph representation learning. In Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, pages 225–234, 2018.

[47] Daniel Oñoro-Rubio, Mathias Niepert, Alberto García-Durán, Roberto González, and Roberto J López-Sastre. Answering visual-relational queries in web-extracted knowledge graphs. arXiv preprint arXiv:1709.02314, 2017.

[48] Xingjia Pan, Fan Tang, Weiming Dong, Yang Gu, Zhichao Song, Yiping Meng, Pengfei Xu, Oliver Deussen, and Changsheng Xu. Self-supervised feature augmentation for large image object detection. IEEE Transactions on Image Processing, pages 6745–6758, 2020.

[49] Heiko Paulheim. Knowledge graph refinement: A survey of approaches and evaluation methods. Semantic web, pages 489–508, 2017.

[50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International Conference on Machine Learning, pages 8748–8763, 2021.

[51] Brandon Reagen, Paul Whatmough, Robert Adolf, Saketh Rama, Hyunkwang Lee, Sae Kyu Lee, José Miguel Hernández-Lobato, Gu-Yeon Wei, and David Brooks. Minerva: Enabling low-power, highly-accurate deep neural network accelerators. In 2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA), pages 267–278, 2016.

[52] Thomas Rebele, Fabian Suchanek, Johannes Hoffart, Joanna Biega, Erdal Kuzey, and Gerhard Weikum. Yago: A multilingual knowledge base from wikipedia, wordnet, and geonames. In International semantic web conference, pages 177–185, 2016.

[53] Feiliang Ren, Juchen Li, Huihui Zhang, Shilei Liu, Bochao Li, Ruicheng Ming, and Yujia Bai. Knowledge graph embedding with atrous convolution and residual learning. arXiv preprint arXiv:2010.12121, 2020.

[54] Hongyu Ren, Weihua Hu, and Jure Leskovec. Query2box: Reasoning over knowledge graphs in vector space using box embeddings. arXiv preprint arXiv:2002.05969, 2020.

[55] Hongyu Ren and Jure Leskovec. Beta embeddings for multi-hop logical reasoning in knowledge graphs. Advances in Neural Information Processing Systems, pages 19716–19726, 2020.

[56] Hongyu Ren and Jure Leskovec. Beta embeddings for multi-hop logical reasoning in knowledge graphs. Advances in Neural Information Processing Systems, pages 19716–19726, 2020.

[57] Andrea Rossi, Denilson Barbosa, Donatella Firmani, Antonio Matinata, and Paolo Merialdo. Knowledge graph embedding for link prediction: A comparative analysis. ACM Transactions on Knowledge Discovery from Data (TKDD), pages 1–49, 2021.

 https://doi.org/10.52202/079017-0306

[58] Dan Ruta, Andrew Gilbert, Pranav Aggarwal, Naveen Marri, Ajinkya Kale, Jo Briggs, Chris Speed, Hailin Jin, Baldo Faieta, Alex Filipkowski, et al. Stylebabel: Artistic style tagging and captioning. arXiv preprint arXiv:2203.05321, 2022.

[59] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. arXiv preprint arXiv:2210.08402, 2022.

[60] Chao Shang, Yun Tang, Jing Huang, Jinbo Bi, Xiaodong He, and Bowen Zhou. End-to-end structure-aware convolutional networks for knowledge base completion. In Proceedings of the AAAI Conference on Artificial Intelligence, pages 3060–3067, 2019.

[61] Chao Shang, Yun Tang, Jing Huang, Jinbo Bi, Xiaodong He, and Bowen Zhou. End-to-end structure-aware convolutional networks for knowledge base completion. In Proceedings of the AAAI Conference on Artificial Intelligence, pages 3060–3067, 2019.

[62] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? arXiv preprint arXiv:2107.06383, 2021.

[63] Ying Shen, Ning Ding, Hai-Tao Zheng, Yaliang Li, and Min Yang. Modeling relation paths for knowledge graph completion. IEEE Transactions on Knowledge and Data Engineering, pages 3607–3617, 2020.

[64] Haoyu Song, Li Dong, Wei-Nan Zhang, Ting Liu, and Furu Wei. Clip models are few-shot learners: Empirical studies on vqa and visual entailment. arXiv preprint arXiv:2203.07190, 2022.

[65] Wenzheng Song, Masanori Suganuma, Xing Liu, Noriyuki Shimobayashi, Daisuke Maruta, and Takayuki Okatani. Matching in the dark: a dataset for matching image pairs of low-light scenes. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 6029–6038, 2021.

[66] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. arXiv preprint arXiv:1908.08530, 2019.

[67] Zequn Sun, Qingheng Zhang, Wei Hu, Chengming Wang, Muhao Chen, Farahnaz Akrami, and Chengkai Li. A benchmarking study of embedding-based entity alignment for knowledge graphs. arXiv preprint arXiv:2003.07743, 2020.

[68] Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. Rotate: Knowledge graph embedding by relational rotation in complex space. arXiv preprint arXiv:1902.10197, 2019.

[69] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. arXiv preprint arXiv:1908.07490, 2019.

[70] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In International conference on machine learning, pages 2071–2080, 2016.

[71] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In NeurIPS, pages 5998–6008, 2017.

[72] Meng Wang, Haofen Wang, Guilin Qi, and Qiushuo Zheng. Richpedia: a large-scale, comprehensive multi-modal knowledge graph. Big Data Research, page 100159, 2020.

[73] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, 2020.

[74] Xin Wang, Benyuan Meng, Hong Chen, Yuan Meng, Ke Lv, and Wenwu Zhu. Tiva-kg: A multimodal knowledge graph with text, image, video and audio. In Proceedings of the 31st ACM International Conference on Multimedia, pages 2391–2399, 2023.

[75] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. Eann: Event adversarial neural networks for multi-modal fake news detection. In Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining, pages 849–857, 2018.

[76] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11686–11695, 2022.

[77] Zikang Wang, Linjing Li, Qiudan Li, and Daniel Zeng. Multimodal data enhanced representation learning for knowledge graphs. In 2019 International Joint Conference on Neural Networks (IJCNN), pages 1–8, 2019.

[78] Fangyun Wei, Yue Gao, Zhirong Wu, Han Hu, and Stephen Lin. Aligning pretraining for detection via object-level contrastive learning. Advances in Neural Information Processing Systems, pages 22682–22694, 2021.

[79] Haoyang Wen, Ying Lin, Tuan Lai, Xiaoman Pan, Sha Li, Xudong Lin, Ben Zhou, Manling Li, Haoyu Wang, Hongming Zhang, et al. Resin: A dockerized schema-guided cross-document cross-lingual cross-media information extraction and event tracking system. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations, pages 133–143, 2021.

[80] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2, 2019.

[81] Ruobing Xie, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. Image-embodied knowledge representation learning. arXiv preprint arXiv:1609.07028, 2016.

[82] Wenhan Xiong, Thien Hoang, and William Yang Wang. Deeppath: A reinforcement learning method for knowledge graph reasoning. arXiv preprint arXiv:1707.06690, 2017.

[83] Derong Xu, Tong Xu, Shiwei Wu, Jingbo Zhou, and Enhong Chen. Relation-enhanced negative sampling for multimodal knowledge graph completion. In Proceedings of the 30th ACM international conference on multimedia, pages 3857–3866, 2022.

[84] Hongwei Xue, Yupan Huang, Bei Liu, Houwen Peng, Jianlong Fu, Houqiang Li, and Jiebo Luo. Probing inter-modality: Visual parsing with self-attention for vision-and-language pre-training. Advances in Neural Information Processing Systems, pages 4514–4528, 2021.

[85] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15671–15680, 2022.

[86] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. In Proceedings of the AAAI Conference on Artificial Intelligence, pages 3081–3089, 2022.

[87] Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. Qa-gnn: Reasoning with language models and knowledge graphs for question answering. arXiv preprint arXiv:2104.06378, 2021.

[88] Zhiwei Zha, Jiaan Wang, Zhixu Li, Xiangru Zhu, Wei Song, and Yanghua Xiao. M2conceptbase: A fine-grained aligned multi-modal conceptual knowledge base. arXiv preprint arXiv:2312.10417, 2023.

[89] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. A large-scale study of representation learning with the visual task adaptation benchmark. arXiv preprint arXiv:1910.04867, 2019.

[90] Jingdan Zhang, Jiaan Wang, Xiaodan Wang, Zhixu Li, and Yanghua Xiao. Aspectmmkg: A multi-modal knowledge graph with aspect-aware entities. In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, pages 3361–3370, 2023.

[91] Ningyu Zhang, Lei Li, Xiang Chen, Xiaozhuan Liang, Shumin Deng, and Huajun Chen. Multimodal analogical reasoning over knowledge graphs. In The Eleventh International Conference on Learning Representations, 2022.

[92] Tongtao Zhang, Ananya Subburathinam, Ge Shi, Lifu Huang, Di Lu, Xiaoman Pan, Manling Li, Boliang Zhang, Qingyun Wang, Spencer Whitehead, et al. Gaia-a multi-media multi-lingual knowledge extraction and hypothesis generation system. In TAC, 2018.

[93] Shangfei Zheng, Weiqing Wang, Jianfeng Qu, Hongzhi Yin, Wei Chen, and Lei Zhao. Mmkgr: Multi-hop multi-modal knowledge graph reasoning. arXiv preprint arXiv:2209.01416, 2022.

[94] Zhehui Zhou, Can Wang, Yan Feng, and Defang Chen. Jointe: Jointly utilizing 1d and 2d convolution for knowledge graph embedding. Knowledge-Based Systems, page 108100, 2022.

[95] Xiangru Zhu, Zhixu Li, Xiaodan Wang, Xueyao Jiang, Penglei Sun, Xuwu Wang, Yanghua Xiao, and Nicholas Jing Yuan. Multi-modal knowledge graph construction and application: A survey. arXiv preprint arXiv:2202.05786, 2022.

 https://doi.org/10.52202/079017-0306

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Sec. D.3
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Sec. D.4
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [NA]
   (b) Did you include complete proofs of all theoretical results? [NA]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Sec. A and Sec. C
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Refer to Tab. 4 for data splits and refer to Sec. C for hyperparameters and other training details.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Sec. C.6.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] In Sec. 4.3, we implement several methods with our *UKnow* , including Q2B [54], BETAE [55], CLIP [50]. Please note that we exclusively utilize the official implementation of Q2B[1] to train our model, while we reimplement the codes for BETAE and CLIP based on the guidelines provided in their respective original papers. Additionally, due credit is given to all creators.
   (b) Did you mention the license of the assets? [NA]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See Sec. A
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [NA]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [NA]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [NA]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [NA]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [NA]

[1] https://github.com/hyren/query2box

A Addition Statement for Our New Dataset

A.1 Dataset Documentation and Intended Use

We offer a detailed overview of our dataset statistics in Sec. 3.3. To facilitate better understanding and ease of access, we have made our dataset project available on ModelScope at: *https://www.modelscope.cn/datasets/yutong/UKnow/summary*, which includes dataset summary, data preview, quickstart and data files.

The detailed data organization and corresponding download links are listed below:

- Original data: We gather our data from publicly available international news sources, accumulating a substantial volume of images and text. Subsequently, we compress the collected data into several zip archives and store them in original_data: `UKnow/raw_data/*`.
- Processed data:
  - Pre-node N_p: Building upon Phase-1, we leverage pre-trained deep learning models to extract valuable information from various domains. The resultant output from Phase-1 is structured as a dictionary and is then stored and saved to pre_node: `UKnow/processed_data/pre_node*`.
  - Node index N_n and Edge index N_e: As the outcomes acquired in Phase-1 (e.g., N_p) are not directly applicable for graph construction, we employ an information symbolization strategy to organize them into indices, namely N_n and N_e, which are subsequently saved to index: `UKnow/processed_data/*_index*.pickle`.
  - Knowledge graph G_m: Finally, we consolidate two types of internal knowledge (I_{in}, T_{in}) and three types of associative knowledge $(I_{cross}, T_{cross}, IT_{cross})$ into into one knowledge graph $(\mathbf{G}_m)$, which is stored as a dictionary in graph: `UKnow/processed_data/graph*.pickle`.

Our dataset is intended for academic use and the corresponding license is based on: https://www.contributor-covenant.org/zh-cn/version/1/4/code-of-conduct.html, which was created by Coraline Ada Ehmke in 2014 and is released under the CC BY-NC-ND 4.0.

A.2 Author statement

We confirm the data licenses and that we bear all responsibility in case of violation of rights.

A.3 Hosting, licensing, and maintenance plan

Hosting and Licensing. Our dataset is hosted on ModelScope. Moreover, we furnish the relevant licenses in accordance with ModelScope at: https://www.contributor-covenant.org/zh-cn/version/1/4/code-of-conduct.html, which was created by Coraline Ada Ehmke in 2014 and is released under the CC BY 4.0 License.

Introduction to ModelScope. ModelScope is a platform designed for managing and optimizing machine learning models. It provides various tools and features to streamline the model development process, including version control, performance monitoring, and collaboration capabilities. As for managing datasets, ModelScope offers robust functionality for organizing, storing, and accessing data. Users can upload datasets to the platform, where they are securely stored and can be easily accessed by authorized team members. ModelScope also supports versioning of datasets, allowing users to track changes over time and ensure reproducibility in their experiments. Additionally, the platform provides tools for data preprocessing, visualization, and analysis, helping users to efficiently prepare their data for model training and evaluation. Overall, ModelScope offers comprehensive support for managing datasets throughout the machine learning lifecycle. Therefore, we choose ModelScope as our hosting platform.

Usage of ModelScope. To enable users to directly utilize all models on the ModelScope platform without configuring the environment, ModelScope integrates an online Notebook programming environment on its website and offers official mirrors for developers. These official mirrors allow users to bypass all installation and configuration steps, providing immediate access to the models. Currently the latest version of the CPU mirror and GPU mirror can be obtained from the office ModelScope repository.

Users also can setup local python environment using following commands:

```
conda create -n modelscope python=3.8
conda activate modelscope
pip install modelscope
```

Then, users can access and enjoy our dataset by:

```
from modelscope.msdatasets import MsDataset
ds =  MsDataset.load('yutong/UKnow', subset_name='default', split='train')
```


https://doi.org/10.52202/079017-0306

Besides, we strongly recommend that users read the official documents for optimal use.

Maintenance Plan. In future work, we will persistently augment the dataset across various scales following the *UKnow* protocol. This endeavor aims to furnish a comprehensive, diverse, and resilient multimodal knowledge graph, thereby facilitating subsequent research endeavors.

B Preliminaries

Multimodal Knowledge Graph. An intuitive interpretation of multimodal knowledge graph is that the ordinary knowledge graph only consists of <head, relation, tail> triples like <("Jony"), Citizen, ("New York")> , but the multimodal knowledge graph consists of the following:
<("Jony"), Citizen, ("NewYork")>,
<("Jony"), Appearance, ("[Face]")>,
<("NewYork"), Landmark, ("[Statueofliberty]")>,
<("[AirForceOne]"), Similarity, ("[AirForceTwo]")>,
where $(\cdot)$ means a text node and $[\cdot]$ means a image node. The machine cannot understand what *"An old man with white hair"* is without establishing the connection between each word and its physical world meaning. However, with the help of multimodal knowledge graph, as a simple example, it is possible to generate a more informative entity-level sentence (*e.g.*, *"Biden is making a speech"*) instead of a vague concept-level description (*e.g.*, *"An old man with white hair is making a speech"*). To evaluate the effectiveness of multimodal knowledge graph (MMKG), several downstream tasks are often performed on the MMKGs, including common-sense reasoning, vision-language pre-training.

Common-sense Reasoning. Common-sense reasoning means answering queries by logic permutations. The specific task in this work is the link prediction. In the inference phase, feeding <("America"), Capital> to a reasoning model, the output should be <("Washington")>. Various works [3, 70, 68, 60, 94, 53] achieve reasoning by embedding entities and relations in knowledge graph into low-dimensional vector space. For instance, GQE [15] encodes queries through a computation graph with relational projection and conjunction ($\wedge$) as operators. Path-based methods [27, 82, 63, 51] start from anchor entities and determine the answer set by traversing the intermediate entities via relational path. There are also GCN [25] based methods [61, 16] pass message to iterate graph representation for reasoning. Common-sense reasoning is an extremely popular task in the field of knowledge graph. Since our dataset is based on the knowledge graph, the performance validation on common-sense reasoning is indispensable.

Vision-Language Pre-training Vision-language pre-training (VLP) can be divided into three categories based on how they encode images [10]: OD-based region features [5, 31, 34, 41, 66, 69], CNN-based grid feature [62, 19, 20] and ViT-based patch features [84, 30, 24]. Pre-training objectives are usually: masked language/image modeling (MLM/MIM) [2, 9, 39], image-text matching (ITM) [34, 19, 10], and image-text contrastive learning (ITC) [30, 50, 35]. In this work, we concentrate on the study of the how to introduce our UKnow into ITC method based on ViT-based patch features.

Image-Text Contrastive Learning. The recent CLIP [50] and ALIGN [21] perform pre-training using a crossmodal contrastive loss on millions of image-text pairs, which achieves remarkable performance on various downstream tasks [42, 62, 64]. MDETR [23] trains on multi-modal datasets which have explicit alignment between phrases and objects. GLIP [32] generates grounding boxes in a self-training fashion, and makes the learned representations semantic-rich. We implement these mainstream methods on our dataset, and also design a basic knowledge-based ITC method with UKnow.

C Experimental Details

In this section, we give more details about the computation complexity, training, fine-tuning hyperparameters and evaluation for reference.

C.1 Common-sense Reasoning

Datasets. Since our dataset is a knowledge graph, we benchmark the performance of KG-reasoning models on our dataset by completing KG-triples. The partitioning of the dataset is illustrated in the upper segment of Tab. 4.

Evaluation. The specific task of common-sense reasoning in this work is the link prediction. Given a test query q (*e.g.*,, <("Jony"), Citizen, (?)>), we are interested in discovering non-trivial answers (*e.g.*,, "New York"). That is, answer entities where at least one edge needs to be imputed in order to create an answer path to that entity. Each entity in our multimodal knowledge graph is not limited to a text entity but a multimodal node. Following [56], for each non-trivial answer t of test query q, we rank it against non-answer entities $\mathcal{E}\backslash[\![q]\!]_{\text{test}}$ [3]. Then the rank of each answer is labeled as r. We use Mean Reciprocal Rank(MRR): $\frac{1}{r}$ and Hits-at-N ($\mathbf{H}@N$) : $1[r \leq N]$ as quantitative metrics.

Table 9: **A new benchmark of the common-sense reasoning task.** We report four metrics of each model on the validation and test sets. All experiments were repeated five times and the variance is shown in the table.

Model	Val-H@1	Val-H@3	Val-H@10	Val-MRR	Test-H@1	Test-H@3	Test-H@10	Test-MRR
TransE [3]	11.75 ± 0.113	29.04 ± 0.112	31.76 ± 0.143	14.77 ± 0.153	11.26 ± 0.114	21.68 ± 0.115	31.57 ± 0.127	14.66 ± 0.123
Q2B [54]	14.99 ± 0.118	25.78 ± 0.135	36.76 ± 0.169	18.80 ± 0.166	14.48 ± 0.119	25.17 ± 0.135	36.32 ± 0.163	18.46 ± 0.134
Q2B*	16.84 ± 0.115	29.00 ± 0.166	38.85 ± 0.169	19.66 ± 0.158	16.35 ± 0.122	28.67 ± 0.174	38.45 ± 0.184	19.27 ± 0.146
BETAE [55]	18.04 ± 0.129	33.02 ± 0.161	41.97 ± 0.179	21.16 ± 0.167	17.65 ± 0.129	32.75 ± 0.160	41.67 ± 0.177	20.75 ± 0.140
BETAE*	19.02 ± 0.125	33.97 ± 0.173	43.17 ± 0.199	21.64 ± 0.173	18.22 ± 0.135	33.52 ± 0.187	42.68 ± 0.198	21.23 ± 0.154
QA-GNN [87]	21.69 ± 0.124	38.11 ± 0.167	45.97 ± 0.180	22.83 ± 0.179	21.05 ± 0.128	37.26 ± 0.164	44.32 ± 0.175	22.06 ± 0.165

Table 10: **A new benchmark of the novel event classification task.** All models are fine-tuned in the training set.

Model	IMG	TXT	Event-11		Event-9185	
			ACC@1	ACC@5	ACC@1	ACC@5
CLIP [50]	✓		65.77	76.82	54.62	63.19
DeCLIP [35]	✓		66.43	78.32	54.86	63.82
ALBEF [30]	✓		66.29	77.84	55.03	63.47
TCL [85]	✓		66.80	78.91	55.87	64.33
CLIP		✓	64.32	75.92	57.48	65.78
DeCLIP		✓	65.89	77.51	59.76	67.81
ALBEF		✓	65.31	76.97	58.43	66.32
TCL		✓	66.03	78.14	59.94	68.23
CLIP	✓	✓	66.08	72.88	57.42	65.65
DeCLIP	✓	✓	67.16	72.96	58.64	66.49
ALBEF	✓	✓	68.03	74.26	60.04	68.13
TCL	✓	✓	68.69	75.02	60.89	69.17

Baselines. We consider four baselines: TransE [3], Q2B [54] and BETAE [56]. Since the *UKnow* based plug-in module can be attached to any reasoning models, we implement the Q2B* with our module based on Q2B and BETAE* based on BETAE. As shown in Tab. 9, BETAE* achieves on average **21.64%** and **21.23%** MRR on the validation and testing set of our dataset, respectively. For a fair comparison (*e.g.*, TransE), our dataset does not construct complex logic such as FOL [14] to evaluate the performance of multi-hop logical reasoning.

C.2 Multimodal Event Classification

We propose a novel task called multimodal event classification, leveraging event annotations (Tab. 3) from both Wiki's event categories and our own manual tagging. The event annotation helps intelligent machines understand human activities and history, offering the possibility to identify which *type of event* or which *real historical event* a picture or a text is relevant to. As shown in Tab. 10, TCL [85] achieves on **66.80%** and **55.87%** on ACC@1 when using the image-input on the *Event-11* and *Event-9185*, respectively. We simply modify all the baseline methods and add a late-fusion module after the image/text encoder to support multimodal classification. Results show that TCL with multimodal inputs obtains gains of **1.89%** and **5.02%** compared with the singlemodal, which demonstrates that multimodal pre-training is more helpful for downstream multimodal tasks.

C.3 Single- & Cross-Modal Retrieval

We design four kinds of single- & cross-modal retrieval tasks: image-to-image, text-to-text, image-to-text, and text-to-image. The construction of GT is based on the event annotations in G_m (Fig. 4). We treat images or texts belonging to the same news event as a similar semantic cluster, and the goal of retrieval is to recall the nearest neighbors within this cluster. The features used for retrieval are derived from the output of the previous layer of the classifier.

As shown in Tab. 11, TCL [85] achieves on **33.24%**, **43.37%** and **45.22%** R@1, R@5, R@10 on the zero-shot setting of image retrieval. The results are **58.89%**, **68.47%** and **73.91%** when fine-tuning the pre-trained parameters, which means the pre-training→fine-tuning strategy is extremely beneficial for downstream retrieval. We provide more details about hyperparameters in Sec. C.5.

C.4 Visual Task Adaptation

Visual Task Adaptation Benchmark (VTAB) [89] is a diverse, realistic, and challenging vision representation benchmark, containing 19 tasks and covering a broad spectrum of domains and semantics. These tasks are grouped into three sets: NATURAL, SPECIALIZED, and STRUCTURED which utilize natural world, professional technology and artificial environment images respectively. We benchmark models on VTAB with ACC@1. We fine-tune models for 10 epoch in each task and compute the inner product between outputs of

https://doi.org/10.52202/079017-0306

Table 11: **A new benchmark of the retrieval task.** Zero-shot means freezing the pre-trained parameters then transfer to the test set for inference. Fine-tune means tuning the pre-trained parameters in the training set before inference.

Model	Retrieval	Zero-Shot			Fine-Tune		
		R@1	R@5	R@10	R@1	R@5	R@10
CLIP [50]	IMAGE	32.41	41.96	43.92	55.97	67.44	71.28
DeCLIP [35]	IMAGE	32.75	42.36	44.38	56.96	66.59	70.95
ALBEF [30]	IMAGE	32.88	42.76	44.79	58.56	67.83	72.24
TCL [85]	IMAGE	33.24	43.37	45.22	58.89	68.47	73.91
CLIP	TEXT	33.02	42.56	46.03	56.50	65.12	70.20
DeCLIP	TEXT	34.00	43.97	47.11	55.87	65.20	70.35
ALBEF	TEXT	33.87	43.86	46.82	56.77	65.91	71.15
TCL	TEXT	34.67	44.25	47.67	56.60	65.50	70.54
CLIP	IMG-to-TXT	32.73	42.64	44.72	56.32	66.93	70.61
DeCLIP	IMG-to-TXT	32.96	42.84	45.17	57.21	66.80	71.26
ALBEF	IMG-to-TXT	33.20	42.97	45.32	58.43	67.59	71.95
TCL	IMG-to-TXT	33.37	43.25	46.04	58.70	67.88	72.33
CLIP	TXT-to-IMG	31.78	41.04	42.51	55.74	64.38	69.56
DeCLIP	TXT-to-IMG	32.13	41.55	42.99	55.84	65.12	70.32
ALBEF	TXT-to-IMG	31.95	41.32	42.85	57.21	66.04	71.50
TCL	TXT-to-IMG	32.56	42.04	43.74	57.17	65.92	71.47

Table 12: **The comparison of *w/* and *w/o UKnow* pre-training.** Zero means the model is initialized with all-zero parameters *w/o* pre-training. CLIP* means pre-training with origin CLIP contrast loss on our dataset. Ours means *UKnow* pre-training.

	CIFAR100	Caltech101	DTD	Flowers102	Pets	SVHN	Sun397	Camelyon	EuroSAT	Resisc45	Retinopathy	ClevrCount	ClevrDist	DMLab	KITTIDist	dSprLoc	dSprOri	sNORBAzim	NORBElev	**VTAB (avg.)**
Zero	58.39	53.54	49.26	52.51	58.93	64.24	48.96	52.44	63.95	60.03	58.62	62.78	62.59	44.27	45.87	75.89	74.48	67.54	60.89	58.69
CLIP*	75.25	71.74	58.39	77.54	74.40	79.42	61.72	70.42	81.56	76.43	67.85	81.25	80.48	60.03	63.98	84.33	82.66	83.68	76.57	74.09
Ours	76.79	72.73	60.44	78.48	76.33	80.56	62.37	72.23	83.27	77.26	65.91	82.46	81.34	63.37	65.74	85.61	82.79	85.12	76.64	**75.23**

images and label texts with prompts [50] through pre-trained image encoders and text encoders as the similarity score. As shown in Tab. 12, our approach obtains gains of avg. **1.14%** compared with the origin CLIP when fairly using the same *UKnow*'s data for the upstream pre-training. For the suboptimal performance on the Retinopathy and NORBElev datasets, we carefully examine the composition of both dataset. The Diabetic Retinopathy dataset consists of image-label pairs with high-resolution retinal images labeled to indicate the presence of diabetic retinopathy (DR) on a scale from 0 to 4. Similarly, the NORBElev dataset contains jittered texture images. It is evident that these data significantly differ from the natural images collected in UKnow. In constract, commonly used general image datasets in practical applications, such as CIFAR-10, tend to show greater improvements when utilizing UKnow. This observation suggests that researchers, when designing advanced knowledge-based pre-training methods with UKnow, should carefully consider balancing data domains according to specific downstream tasks. Additionally, accurate node construction is essential for building a robust multimodal knowledge graph to fully leverage the advantages of UKnow. This underscores the importance of designing effective pre-processing functions P, particularly in specialized subfields such as the Retinopathy dataset. In these domains, more dedicated data pre-processing models, such as medical image segmentation and detection models, can be employed to enhance feature extraction.

The backbone of CLIP is ViT-B/32. The cost of pre-train is 26h / 30epoch. The key hyperparameters are *bs: 512, lr: 0.001, warmup: 1e4, eps: 1e-8, beta1: 0.9, beta2: 0.999, dim: 512, AdamW*. The detailed setting can be found in Sec. C.5. It is essential to highlight that the image-text PAIR constitutes only one type of data in our protocol. By leveraging the capabilities of *UKnow* , our pre-trained CLIP model can effectively comprehend the inherent knowledge ingrained within the data, resulting in superior performance than the original CLIP model (as observed in Tab. 12, Row2, utilizing image-text PAIR only).

C.5 Hyperparameters

Tab. 13 and Tab. 14 list the hyperparameters that differ on each models and are determined with the validation performance on our dataset. In particular, Tab. 13 lists 7 common hyperparameters, such as learning rate, batch size, warmup, epoch number, *etc.*, employed during pre-training. The pre-trained model is evaluated using a standard pipeline consisting of pre-training on Dataset1, fine-tuning on Dataset2-Train, and testing on either Dataset2-Test/Val. Therefore, we list the hyperparameters used during fine-tuning in Tab. 14, which are slightly

different from Tab. 13. We omit some of the model results, since ALBEF and TCL share the same set of hyperparameters, and the original CLIP and CLIP-UKnow share the same set of parameters.

Table 13: **Hyperparameters for models of pre-training.**

Hyperparameter	ALBEF	DeCLIP	CLIP-UKnow
Learning Rate	0.0001	0.001	0.001
Batch Size	128	128	512
Number of Epochs	30	30	30
Weight Decay	0.02	0.1	0.1
Optimizer	`AdamW`	`AdamW`	`AdamW`
Feature Dim	256	512	512
Warmup	20epc	5000	10000

Table 14: **Hyperparameters for models of fine-tuning.**

Hyperparameter	ALBEF	DeCLIP	CLIP-UKnow
Learning Rate	0.0001	5e-5	5e-5
Batch Size	128	256	256
Number of Epochs	128	20	20
Weight Decay	0.02	0.02	0.02
Optimizer	`AdamW`	`AdamW`	`AdamW`
Feature Dim	256	512	512
Warmup	4epc	6epc	6epc

C.6 Computation Complexity

Here we detail the time cost of pre-training and fine-tuning. The GPU is NVIDIA(R) A100, the memory of GPU is 81,251MiB, driver version is 470.154, CUDA version is 11.4. The CPU is Intel(R) Xeon(R) Platinum 8369B @ 2.90GHz with 15 physical computation cores. The environment is Python 3.6.12 with Torch 1.10.1. Results are as shown in Tab. 15 and Tab. 16.

Table 15: **The time cost of pre-training.**

Model	Backbone	Epoch	Batch	Time/h
DeCLIP	ViT-B/32	30	128	91
ALBEF	ViT-B/16	30	128	69
TCL	ViT-B/16	30	128	67
CLIP*	ViT-B/32	30	512	25
CLIP-UKnow	ViT-B/32	30	512	26

Table 16: **The time cost of downstream fine-tuning.**

Model	Backbone	*UKnow* Tasks			VTAB		
		Epoch	Batch	Time/h	Epoch	Batch	Time/h
DeCLIP	ViT-B/32	20	128	12	-	-	-
ALBEF	ViT-B/16	20	128	10	-	-	-
TCL	ViT-B/16	20	128	10	-	-	-
Zero*	ViT-B/32	-	-	-	15	128	3
CLIP*	ViT-B/32	20	256	8	15	128	3
CLIP-UKnow	ViT-B/32	20	256	8	15	128	3

D Discussion

D.1 Complexity

We notice that the detailed pipeline and protocol may appear complex and require effort to implement and understand fully. However, this complexity is necessary to ensure that the pipeline is robust, flexible, and capable of handling diverse and multimodal datasets.

To mitigate the implementation challenges, we have designed the pipeline to be modular, like Phase-1/2/3, allowing each phase to be independently replaced, added, or disabled based on specific needs. Moreover, we present an extra dataset documentation and construct a website in Sec. A.1. It provides a detailed data organization, corresponding download links, and an example code to guide users through the process, making the protocol more accessible and easier to adopt. Our goal is to balance complexity with practicality, ensuring that the benefits of a thorough and versatile approach outweigh the initial learning curve.

https://doi.org/10.52202/079017-0306

D.2 Correlation between the Knowledge View and Phase 1

In Phase 1, Content Extraction is designed to preprocess raw data (such as images and texts) using pre-trained deep learning models, which extract essential information that serves as the foundation for our knowledge view. The extracted content N_p provides a rich, structured collection of attributes and features that capture both global and semantic-level details from the input. It transforms raw data into a set of key-value pairs that represent various aspects of the input content. These key-value pairs encapsulate knowledge at different levels, which are critical for constructing meaningful nodes in the subsequent phases. This structured output essentially forms the knowledge view of our system, where each extracted piece of information is treated as a node attribute. These attributes are later symbolized and linked in Phase 2, leading to the construction of the multimodal knowledge graph in Phase 3. Thus, the content extracted in Phase 1 is directly correlated with the knowledge view, serving as the core data that the entire graph construction process relies upon.

D.3 Limitation and Future Work

Despite the strides made, our research bears certain limitations. First of all, our current dataset primarily centers on text and image modalities which serve as fundamental pillars for information storage and representation, but lack other useful modalities. In future work, we aim to diversify modalities by augmenting our dataset with a broader range of modalities (*e.g.,* audio, video, 3D, etc.) to facilitate exploration across various downstream tasks. Second, for each downstream task, we selected several basic yet most suitable methods for our work as our baseline, resulting in slight deviations with current state-of-the-art (SOTA) performance. Our primary objective lies in validating the efficacy of our proposed dataset and protocols, and demonstrating the most straightforward and intuitive approach for utilizing our dataset. Hence, we made certain trade-offs, sacrificing some performance by opting for a more rudimentary approach instead of pursuing the SOTA method to enhance understanding and usage. We anticipate that our simplified demonstration will stimulate the community to delve deeper into the potential enhancements that *UKnow* can offer in improving performance.

D.4 Societal Impact

As stated in Sec. 3.2, our dataset originates from publicly accessible international news sources via the Wikipedia API. These sources only contain events that are publicly available and do not include any sensitive information. Consequently, we confidently affirm that our research carries no potential negative societal impacts.