# TIVA-KG: A Multimodal Knowledge Graph with Text, Image, Video and Audio

Xin Wang
Department of Computer Science and Technology, BNRist, Tsinghua University
xin_wang@tsinghua.edu.cn

Benyuan Meng
Department of Computer Science and Technology, Tsinghua University
mby18@mails.tsinghua.edu.cn

Hong Chen
Department of Computer Science and Technology, Tsinghua University
h-chen20@mails.tsinghua.edu.cn

Yuan Meng
Department of Computer Science and Technology, Tsinghua University
meng-y16@mails.tsinghua.edu.cn

Ke Lv
The University of the Chinese Academy of Sciences
luk@ucas.ac.cn

Wenwu Zhu*
Department of Computer Science and Technology, BNRist, Tsinghua University
wwzhu@tsinghua.edu.cn

## ABSTRACT

Knowledge graphs serve as a powerful tool to boost model performances for various applications covering computer vision, natural language processing, multimedia data mining, etc. The process of knowledge acquisition for human is multimodal in essence, covering text, image, video and audio modalities. However, existing multimodal knowledge graphs fail to cover all these four elements simultaneously, severely limiting their expressive powers in performance improvement for downstream tasks. In this paper, we propose TIVA-KG, a multimodal Knowledge Graph covering Text, Image, Video and Audio, which can benefit various downstream tasks. Our proposed TIVA-KG has two significant advantages over existing knowledge graphs in i) coverage of up to four modalities including text, image, video, audio, and ii) capability of triplet grounding which grounds multimodal relations to triples instead of entities. We further design a Quadruple Embedding Baseline (QEB) model to validate the necessity and efficacy of considering four modalities in KG. We conduct extensive experiments to test the proposed TIVA-KG with various knowledge graph representation approaches over link prediction task, demonstrating the benefits and necessity of introducing multiple modalities and triplet grounding. TIVA-KG is expected to promote further research on mining multimodal knowledge graph as well as the relevant downstream tasks in the community. TIVA-KG is now available at our website: http://mn.cs.tsinghua.edu.cn/tivakg.

## CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**.

*Corresponding author. BNRist is the abbreviation of Beijing National Research Center for Information Science and Technology.



*MM '23, October 29-November 3, 2023, Ottawa, ON, Canada*

ACM ISBN 979-8-4007-0108-5/23/10.
https://doi.org/10.1145/3581783.3612266

## KEYWORDS

Knowledge Graph, Multimodal, Text, Video, Image, Audio

**ACM Reference Format:**
Xin Wang, Benyuan Meng, Hong Chen, Yuan Meng, Ke Lv, and Wenwu Zhu. 2023. TIVA-KG: A Multimodal Knowledge Graph with Text, Image, Video and Audio. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3581783.3612266

## 1 INTRODUCTION

Knowledge graph (KG) is an effective way to explicitly store and utilize knowledge, which supports and boosts model performances in various domains ranging from computer vision, natural language processing and multimedia analysis. Typically, KG encodes knowledge in the form of triples *<head, relation, tail>*, forming a multi-relation heterogeneous graph. In this paper, "triple" is interchangeably used with "triplet". With the increasing amount of multimodal data becoming publicly available for various multimedia tasks, multimodal knowledge graph (MMKG), i.e., KG with multimodal information associated with nodes, has attracted more and more attention from the research community. There have been a few works that utilize MMKG as external knowledge sources for multimodal tasks, such as Richpedia [35], MMKG [19] and VisualSem [1]. This is consistent with the process of knowledge acquisition for human, which is multimodal in essence covering text, image, video and audio.

However, there exist two major weaknesses in the current MMKG works.

- Existing works on MMKG only cover at most two modalities simultaneously, mostly covering text and image, other work such as WASABI [4] contains audio and text, and VideoGraph [26] contains video and text. These works fail to cover all four elements of text, image, video and audio simultaneously, severely limiting their expressive powers in performance improvement for downstream tasks.
- Whilst multiple entities and relations can be combined to express a complex symbolic concept, multimodal data grounded to them cannot be naturally combined. In order to find suitable multimodal knowledge for such a complex symbolic

concept, triplet grounding is beneficial [41], which grounds multimodal data to whole triples instead of single entities. For example, *a dog is able to bark* should ideally be characterized by a triple of entities <Dog, IsAbleTo, Bark> as a whole to reflect the symbolic knowledge, rather than being characterized through three separate entities representing Dog, and Bark independently. Nevertheless, the capability of triplet grounding has been largely ignored by existing works.

To tackle these issues, in this paper we propose TIVA-KG, a multimodal Knowledge Graph covering Text, Image, Video and Audio simultaneously, as well as providing the capability of triplet grounding. To the best of our knowledge, TIVA-KG is the first general KG simultaneously including text, image, video and audio modalities together. With the novel design of associating multimodal attributes with both entities and triples, our proposed TIVA-KG is able to conduct triplet grounding that captures symbolic knowledge carried in KG, e.g., entity(Dog)–relation(IsAbleTo)→ entity(Bark). Our design of triplet grounding is able to boost the ability of expressing both specific and complicated concepts when utilizing multimodal information of KG. Take another triple <Dog, CapableOf, Run> illustrated in Figure 1 as an example, i) entity Dog is characterized by multimodal data which demonstrate dogs sitting or standing, ii) entity Run is characterized with multimodal data describing the scenario of human running, and iii) triplet <Dog,CapableOf, Run> is grounded via multimodal data indicating the running dogs.

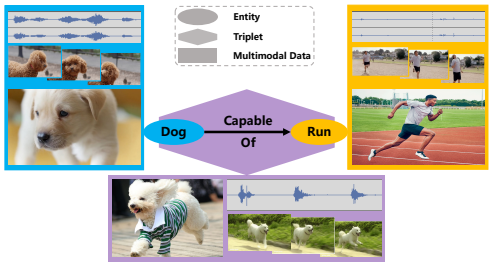


**Figure 1: An example of entity grounding for Dog (blue), entity grounding for Run (yellow) and triplet grounding for <Dog, CapableOf, Run> (purple).**

To construct TIVA-KG, we first extract a subgraph from ConceptNet [32] focusing on general knowledge, which serves as the initial skeleton of TIVA-KG. Next, we build up an automatic crawler to acquire data of image, video and audio modalities through caption-based approach [41] which generates a natural language description for each entity and triplet to search from Google and FreeSound. The data crawled from the web can be further processed into feature vectors for subsequent analysis over TIVA-KG.

Besides, we design a Quadruple Embedding Baseline (QEB) model to integrate information from text, image, video and audio modalities as well as triplet grounding for link prediction on KG. We conduct extensive experiments through comparing both existing unimodal and multimodal approaches with our QEB model, as well as benchmarking the link prediction task on our TIVA-KG. Experimental results show significant performance increase of QEB on TIVA-KG and demonstrate the importance of both two novel features of TIVA-KG.

**Table 1: Comparison between TIVA-KG and other public multimodal knowledge graphs (MMKGs)**

| MMKG | modality | | | | multimodal knowledge |
|---|---|---|---|---|---|
| | text | image | video | audio | |
| IMGpedia | ✓ | ✓ | ✗ | ✗ | entity, relation |
| ImageGraph | ✓ | ✓ | ✗ | ✗ | entity |
| MMKG | ✓ | ✓ | ✗ | ✗ | entity |
| Richpedia | ✓ | ✓ | ✗ | ✗ | entity, relation |
| VisualSem | ✓ | ✓ | ✗ | ✗ | entity |
| TIVA-KG | ✓ | ✓ | ✓ | ✓ | entity, **triplet** |

In summary, this work makes the following contributions:

- We introduce TIVA-KG, a new large-scale multimodal KG containing texts, images, videos and audio together. To the best of our knowledge, TIVA-KG is the first general KG that covers four modalities simultaneously.
- We propose *triplet grounding* on multimodal KG, which is able to ground symbolic knowledge on TIVA-KG, thus significantly boosting the expressiveness of knowledge representation over KG with multimodal information.
- We design Quadruple Embedding Baseline (QEB), a new baseline model to exploit text, image, video and audio modalities simultaneously for multimodal knowledge representation over TIVA-KG.
- We conduct extensive experiments on TIVA-KG and compare our QEB with several state-of-the-art approaches ranging from unimodal to bimodal setting, demonstrating the advantages of QEB over existing methods as well as the necessity of quadruple modalities and triplet grounding.

## 2 RELATED WORKS

**Existing Multimodal Knowledge Graphs.** IMGpedia [6] is one of the first attempts to collect images and form a KG, containing only image modality. MMKG [19] follows a more traditional philosophy, enriching DBPEDIA, YAGO and Freebase-15k with numeric literals and image information to form an MMKG, which also provides an early example of typical practice for MMKG construction. Whereas Richpedia [35] pays more attention to improve data quality and filter images through a distinctive retrieval model, VisualSem [1], as a more recent work, simultaneously builds a novel image filtering pipeline and provides multimodal retrieval models that retrieve entities given images and sentences. With all these topics explored, however, more attention can still be paid to the combination of more (e.g., quadruple) modalities [39] together and triplet grounding. Table 1 shows a comparison between our proposed TIVA-KG and existing MMKGs.

**Link Prediction on KG.** MMKGs serve as a knowledge base for a wide range of downstream tasks. These tasks can be classified into two categories: in-KG tasks and out-of-KG tasks [17, 29], depending on whether they require additional labeled data or not [41]. In-KG tasks refer to tasks that are conducted entirely within the scope of the MMKG, and there are three primary types: knowledge graph

completion (KGC), relation discovery and entity discovery [12]. KGC aims to expand existing KGs by predicting new links between entities based on the available information in the MMKG. Relation discovery and entity discovery, on the other hand, are focused on extracting new knowledge from text. To evaluate the quality of TIVA-KG, we focus on the performance of MMKG on link prediction tasks.

In recent years, there has been a surge of research in deep learning-based approaches for link prediction, which learn low-dimensional embeddings to represent entities and relations. One common approach is to define a scoring function $\phi(\mathbf{h}, \mathbf{r}, \mathbf{t})$ to estimate the plausibility of a given fact using the embeddings of its entities and relations [27].

Some models use tensor decomposition to learn these embeddings, with DistMult [37] and ComplEx [34] being popular examples. These models force relation embeddings to be diagonal matrices, which reduces the number of parameters and makes them easier to train. SimplE [13] also uses diagonal relation embeddings, but can model asymmetric relations by incorporating inverse-direction information. Analogy [18] adds constraints on a general bilinear scoring function, inspired by analogical structures. While HolE [23] computes circular correlation between head and tail entity embeddings to reduce time and space complexity. TuckER [2] uses the Tucker decomposition [14] to factorize a tensor into a set of vectors and a shared core. Geometric models utilize geometric transformations in the latent space to interpret relations, with TransE [3] being a popular example. TransE requires the tail embedding to lie close to the sum of the head and relation embeddings, but suffers from limitations on handling one-to-many, many-to-one and many-to-many relations. STransE [22] pre-multiplies head and tail embeddings with relation-specific matrices to address these limitations. CrossE [40] combines element-wise products with triple-specific embeddings, while RotatE [33] allows for modeling relational patterns such as symmetry/anti-symmetry, inversion, and composition through rotations in a complex latent space. TorusE [5] projects points from the Euclidean space onto a torus to handle the translational constraint of TransE.

In medicine related research fields, interpretability of link prediction results is critical, and rule-based methods have received attention. Rule mining algorithms [7, 8, 15, 20] often rely on preset metrics like confidence and support, but suffer from limitations in relying on discrete counting. Neural-LP [38] and DRUM [28] combine parameter and structure learning of first-order logical rules in an end-to-end differentiable model. RNNLogic [25] treats logic rules as a latent variable and trains a rule generator and reasoning predictor simultaneously, under the EM framework.

Furthermore, there exists work [21] which tries to design models capable of utilizing multimodal knowledge to get better performance on MMKGs. But there have been no existing models that can directly utilize triplet multimodal knowledge, motivating us to propose a new baseline method capable of tackling this problem.

# 3 TIVA-KG: KNOWLEDGE GRAPH WITH TEXT, IMAGE, VIDEO AND AUDIO

In this section, we conceptually discuss the establishment of TIVA-KG with respect to sources and ontology. TIVA-KG focuses on general knowledge, e.g., animals, social relationships and geology etc., which is gathered from multiple sources and represented via entities and triplets. These entities and triplets in TIVA-KG will be aligned with multimodal information through our construction. We also present the detailed statistics of the proposed TIVA-KG and visualize its subgraph with 30K entities.

## 3.1 Sources

The knowledge carried in TIVA-KG contains two types of information,

(1) Structural and textual information associated with the basic topology.
(2) Image, audio and video information associated with entities and triplets.

The basic topology of TIVA-KG is extracted from ConceptNet [32], a publicly available single modality knowledge graph carrying general knowledge via texts, which is gathered from multiple sources. As such, by inheriting the fundamental topology from ConceptNet, TIVA-KG naturally benefits in the same quality and diversity in terms of structural and textual information with ConceptNet.

In addition, TIVA-KG further incorporate multimodal data covering images, videos and audio from Google and Freesound through a web crawler specifically designed for TIVA-KG. Given a natural language description, our web crawler is able to retrieve multimodal information by utilizing search engines of Google and Freesound. The retrieved results from the search engines are ranked, and the top results will be picked up to guarantee that TIVA-KG receives highly relevant results with high possibility.

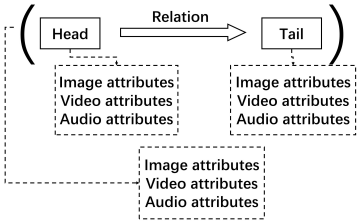


**Figure 2: TIVA-KG ontology**

## 3.2 Ontology

Upon inheriting the advantages of ConceptNet, our proposed TIVA-KG is able to further carry information from image, video and audio modalities, as well as being capable of direct grounding on triplet <entity, relation, entity>.

**Handling Topological Structure.** The basic topology of ConceptNet can be regarded as a multi-relational graph, where nodes indicate entities representing different concepts such as "Cat", "Pet" etc. and edges indicate relations such as "IsA" and "UsedFor" etc. By combining two entities as well as the relation between them together, it is possible to form a triplet capable of providing more expressive information than separate entities. For instance, by combining entity "Cat", entity "Pet" and relation "UsedFor" together, we can get triplet "Cat–UsedFor→Pet". Entities, relations and triplets are common elements shared across all types of KGs to date, which are also adopted by TIVA-KG.

**Handling Multimodal Information.** As for the ontology regarding multimodal data, existing MMKGs such as Richpedia [35] use different types of nodes to represent multimodal data, and utilize different types of edges to connect different types of nodes. For example, a relation of type "ImageOf" may originate from an *image node* to an *entity node* while a relation of type "ImageSimilarity" can connect two *image nodes*. However, this design fails to conduct triplet grounding with multimodal information.

To enable triplet grounding with multimodal information, we organize multimodal information associated with each node or triplet as attributes in TIVA-KG. In concrete, multimodal data associated with entities will be stored as entity attributes, and multimodal data associated with triplets will be stored as attributes of triples. Therefore, our design for storing multimodal information in TIVA-KG is able to concisely represent relational knowledge carried within triplets in a natural and concise way. Figure 2 shows the basic ontology of TIVA-KG.

## 3.3 Statistics and Visualizations

TIVA-KG consists of 440K entities and 1.3M triples, i.e., 443,580 entities and 1,382,358 triples, with every entity reachable from others to ensure good connectivity. In TIVA-KG, multimodal data can associate with both entities and triplets, where each modality has at most 5 data samples to be stored. Table 2, Table 3, Table 4 provide a detailed statistics for entities, triplets and top-10 entities of the largest degree. Figure 3 demonstrates the percentage of each relation type in TIVA-KG.

**Table 2: Statistics of entities in TIVA-KG.**

| | # entities covering the corresponding modality | # data samples in each modality |
|---|---|---|
| Audio | 103,580 | 359,465 |
| Image | 340,225 | 1,695,688 |
| Video | 239,566 | 1,112,918 |

**Table 3: Statistics of triplets in TIVA-KG.**

| | # triplets covering the corresponding modality | # data samples in each modality |
|---|---|---|
| Audio | 30,169 | 93,521 |
| Image | 223,998 | 1,117,389 |
| Video | 194,037 | 927,029 |

Figure 4 provides a visualization of subgraph with 30K entities extracted from TIVA-KG.

# 4 CONSTRUCTION, STORAGE AND ACCESSIBILITY

In this section, we explain the detailed process of constructing, storing and accessing TIVA-KG to ease the utilization of TIVA-KG in various tasks.

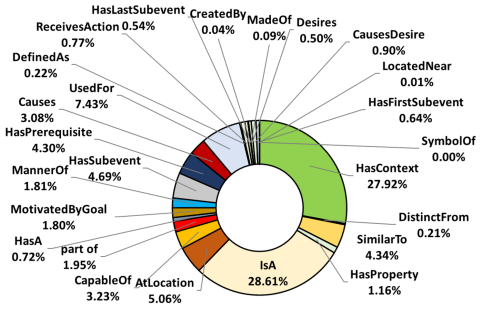


**Figure 3: Percentage of each relation type in TIVA-KG.**

**Table 4: Top-10 entities of the largest degree.**

| Entity | Degree | Entity | Degree |
|---|---|---|---|
| slang | 14,137 | organic compound | 9282 |
| us | 11,452 | chemistry | 9258 |
| zoology | 10,332 | archaic | 9196 |
| medicine | 10,092 | computing | 8807 |
| historical | 9660 | uk | 8141 |

## 4.1 Constructing TIVA-KG

The constructing procedure for TIVA-KG mainly consists of three steps: i) We extract a skeleton from ConceptNet as the basic topology; ii) we associate multimodal data to entities and triplets within the basic topology; iii) We transform the raw multimodal data into latent features in vector form.

*4.1.1 Basic Topology.* We conduct a filtering procedure over ConceptNet to obtain the basic topology suitable for associating multimodal information with entities and triplets, based on the following rules.

(1) We conduct filtering based on language tags, only including English entities as well as English triplets, and excluding externalURL relation.
(2) Since there are weight attributes in ConceptNet, we improve the data quality via dropping relations with weight smaller than 1, which are usually noises or wasted data in the original ConceptNet.
(3) We remove relations of certain types with ambiguous semantic meaning, e.g., RelatedTo. Relation types that are KEPT in TIVA-KG include IsA, PartOf, HasA, UsedFor, CapableOf, HasProperty, MannerOf, MadeOf, ReceivesAction, AtLocation, Causes, HasSubevent, HasFirstSubevent, HasLastSubevent, HasPrerequisite, MotivatedByGoal, ObstructedBy, Desires, CreatedBy, DistinctFrom, SymbolOf, DefinedAs, LocatedNear, HasContext, SimilarTo, CausesDesire.

We conduct Breadth-First Search (BFS) to apply the above filtering rules simultaneously, starting from the node "cat" and stopping when no new neighbors are discovered by BFS anymore. During the filtering process, those excluded entities and relations will be ignored.

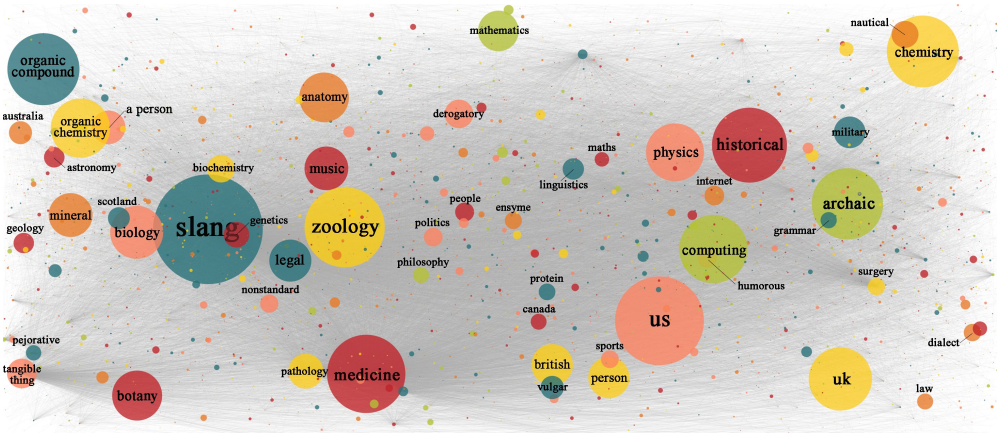


**Figure 4: Visualization of subgraph containing 30K entities extracted from TIVA-KG.**

*4.1.2 Association of Multimodal Data.* We adopt the caption-based approach [41] to conduct entity grounding and triplet grounding, which is also used in the construction of some other MMKGs [19, 24]. We pick one entity or triple, generate its natural language description, and then use the description to search images, GIF files, audio clips on Google and Freesound. Given that some types of relations aim to provide structural information, lacking explicit semantic meaning suitable for associating multimodal data, we ignore the following relation types when aligning multimodal data: IsA, MannerOf, HasSubevent, HasFirstSubevent, HasLastSubevent, HasPrerequisite, MotivatedByGoal, ObstructedBy, Desires, DistinctFrom, SymbolOf, DefinedAs, HasContext, SimilarTo, CausesDesire, NotDesires.

We employ the tool provided by ConceptNet to generate "labels" for each entity, which can be used as the natural language descriptions related to the corresponding entity directly. By further combining relation type and the natural language descriptions of its related entities, we are able to generate a textual description for each triplet. For instance, the detailed rules to generate the descriptions for "A–r→B" are as follows:

- "A B" if r is PartOf, HasProperty, AtLocation, or Cause.
- "A has B" if r is HasA.
- "A used for B" if r is UsedFor.
- "A can B" if r is CapableOf.
- "A is made of B" if r is MadeOf.
- "A can be B" if r is ReceivesAction.
- "A created by B" if r is CreatedBy.
- "A near B" if r is LocatedNear.

With the generated descriptions, we search data from other modalities through various ways: i) For image modality, we search on Google; ii) For video modality, we search on Google and specify the data type as *.gif*; iii) For audio modality, we search on FreeSound. For each modality of every single entity or triplet, at most 5 data samples are retrieved, whose orders are determined according to the ranking from Google or FreeSound. We observe that the textual descriptions are generated based on semantic information of the corresponding entities and triplets, therefore the resulting multimodal data will be naturally aligned together with no need for further processing.

*4.1.3 Latent Features.* We provide preprocessed features in vector form instead of raw multimodal data for the sake of copyright issue and efficient storage at the time of writing. The features are extracted and processed as follows.

- For text features, we provide word embeddings (i.e., semantic vectors) for entities, which are inherited from ConceptNet. These vectors, called *ConceptNet Numberbatch* [31], are trained via combining textual and structural information in ConceptNet, being able to provide vectors of the same length no matter how many words an entity may actually contain. We also keep all the original texts so that alternative methods can still be used to extract textual features.
- For image features, we employ ResNet-101 [9] to obtain a 2048-dimension feature vector for every data sample within each entity or triplet.
- For audio features, we adopt VGGish [10] for the processing procedure where the raw audio is resampled and key frames are chosen within certain intervals, resulting latent factors of shape (x, 128) with x depending on raw audio duration time.
- For video features, we utilize HCRN [16] for video feature processing. First, the video is sampled into 8 clips with identical intervals, each of which containing 16 continuous frames. Then these frames are separately fed into ResNet-101 to obtain frame features, and 8 clips together are fed into ResnNet-101 to get motion features.

In the end, we provide an example to better illustrate the construction procedure. Taking the triple <Dog, CapableOf, Run> as an example, it exists in the original ConceptNet. When we obtain the basic topology from ConceptNet, it passes all filtering rules and thus is included in TIVA-KG. Then we use a caption-based approach to align multimodal knowledge to it. The generated natural language phrase is "dog can run", and this phrase is used to search for images as well as videos on Google, and audio clips on FreeSound. Finally, the obtained multimodal files are processed into latent features.

## 4.2 Storing and Accessing TIVA-KG

**Storing TIVA-KG.** The basic topology and multimodal data of TIVA-KG are stored in an independent manner. The Basic topology

can be used alone to regard TIVA-KG as a single modality general KG, as well as be used jointly with multimodal data as a multimodal general KG with four modalities.

TIVA-KG adopts a new way of representing topology for structural information and storing multimodal attribute (such as URI link) for multimodal data. In concrete, we assign each entity or triplet a unique ID and store entities and triplets in two separate dictionaries, i.e., *entity dictionary* and *triplet dictionary*, which are accessible through IDs. In the *entity dictionary*, each entity entry records multimodal attributes as well as the IDs of triplets relevant to this entity. In the *triplet dictionary*, each triplet entry records multimodal attributes and the IDs of both entities in the corresponding triplet. The recorded multimodal attributes are actually URI links which can direct to the real corresponding multimodal data. We remark that separating the storage of entity and triplet information into two separate dictionaries may help to avoid storing redundant information. Both of the *entity dictionary* and *triplet dictionary* data are organized into single JSON files, which are straightforward to use.

Following the URI links, one can reach TIVA-KG's multimodal data. Raw multimodal data is stored in the file system individually, while latent features are organized into one HDF5 file, and both share the same URI.

**Accessing TIVA-KG.** The files containing dictionaries and other additional information such as structural embedding features and multimodal features are now available online. It is necessary to mention that many existing works represent the topology of KG as Resource Description Framework (RDF) triplets, where a KG is usually stored in a *triple file*. We find it simple to transform TIVA-KG into *triple files* to be compatible with such prior codes. To do so, we can traverse the triplet dictionary of TIVA-KG, and convert each entry into a line in the *triple file*. To keep track of multimodal information for the triplet, it is necessary to add an extra column in the *triple file* so that the original triplet IDs are still accessible. This transformation makes it easy to adapt existing codes to TIVA-KG.

## 5 QUADRUPLE EMBEDDING BASELINE

In this section, we discuss in detail the proposed QEB, a **Q**uadruple **E**mbedding **B**aseline model which is able to fully exploit multimodal knowledge of both entities and triplets.

### 5.1 Energy Functions

We denote a KG as $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T})$, where $\mathcal{E}$ is the set of all entities, $\mathcal{R}$ is the set of all relations, and $\mathcal{T} = \{(h, r, t)|h, t \in \mathcal{E}, r \in \mathcal{R}\}$ is the set of all triplets.

We adopt the common practice of translation models, e.g., TransE [3]. For a triplet $(h, r, t)$, we denote feature vectors related to *head*, *relation* and *tail* as $h$, $r$ and $t$, respectively, satisfying the translational assumption $h + r \approx t$. This denotation is simple yet effective, capable of being implemented in many different ways via replacing the three feature vectors with more alternative ones.

We define two types of embeddings for entities and relations: structural embeddings $h_s^I, r_s^I, t_s^I \in \mathbb{R}^N$ directly obtained from TransE [3], and multimodal embeddings $h_m^I, t_m^I \in \mathbb{R}^{M_1}, r_m^I \in \mathbb{R}^{M_2}$, where $I$ refers to the input embeddings. Given that the embeddings come from different spaces, we project them into a common latent space through a multi-layer network, obtaining $h_s, r_s, t_s, h_m, r_m, t_m \in \mathbb{R}^P$, indicating structural representation of *head* ($h_s$), *relation* ($r_s$) and *tail* ($t_s$) as well as the multimodal representation of *head* ($h_m$), *relation* ($r_m$) and *tail* ($t_m$). Furthermore, given a triplet $(h, r, t)$, we follow the common practice [21] to define three groups of energy functions, i.e., i) Intra-Embedding Energy, ii) Inter-Embedding Energy and iii) Complementary Energy.

**Intra-Embedding Energy.** By extending the structural energy defined by the TransE approach, we define the intra-embedding energy via calculating the distance between embedding vectors obtained from either structural or multimodal information as follows,

$$E_s = ||h_s + r_s - t_s||, \quad E_m = ||h_m + r_m - t_m||.$$

**Inter-Embedding Energy.** Although the structural and multimodal input embeddings are required to share the same number of dimensions, they are not guaranteed to share the same embedding space. As such, we further define the inter-embedding energy, through the six possible combinations across structural and multimodal embedding space as follows,

$$E_{MSM} = ||h_m + r_s - t_m||,\ E_{MSS} = ||h_m + r_s - t_s||,\ E_{SSM} = ||h_s + r_s - t_m||,$$
$$E_{SMS} = ||h_s + r_m - t_s||,\ E_{SMM} = ||h_s + r_m - t_m||,\ E_{MMS} = ||h_m + r_m - t_s||.$$

These functions indicate i) the relation corresponding to a translation operation between the multimodal (structural) representation of the head and tail entities once projected into the structural (multimodal) space (i.e., *MSM* and *SMS*); and ii) the constraint [36] of ensuring the structural and the multimodal representations to be learned in the same space (*MSS, SSM, SMM, MMS*).

**Complementary Energy.** Besides $E_{MSM}$ and $E_{SMS}$ of Inter-Embedding Energy, we enforce the constraint additionally on the summation of multimodal and structural embeddings as the complementary energy to improve robustness as follows,

$$E_{CS} = ||(h_m + h_s) + r_s - (t_m + t_s)||,$$
$$E_{CM} = ||(h_m + h_s) + r_m - (t_m + t_s)||.$$

**Putting All Together.** The overall energy for a triplet with two end nodes (i.e., *head* $h$ and *tail* $t$) and one *relation* $r$ can be defined as the sum of intra-embedding energy, inter-embedding energy and complementary energy in the following,

$$E(h, r, t) = E_s + E_m + E_{CS} + E_{CM} + E_{MSM} + E_{SMS} + E_{MSS} + E_{SSM} + E_{SMM}, + E_{MMS}. \quad (1)$$

### 5.2 Objective Function

Following the common practice [21], the model is trained to ensure that the overall energy of positive sample $E(h, r, t)$ or $E(t, -r, h)$ ($-r$ refers to the reversed relation) is minimized while the overall energy of negative sample $E(h, r, t')$ or $E(t, -r, h')$ ($t'$ and $h'$ refer to negative *tail* node and *head* node respectively) is maximized through a margin-based ranking loss between the overall energies of positive and negative samples.

$$L_{\text{head}} = \sum_{(h,r,t)\in T} \sum_{(h,r,t')\in T'_{\text{tail}}} max(\gamma + E(h, r, t) - E(h, r, t'), 0),$$
$$L_{\text{tail}} = \sum_{(h,r,t)\in T} \sum_{(h',r,t)\in T'_{\text{head}}} max(\gamma + E(t, -r, h) - E(t, -r, h'), 0),$$

where $\gamma$ serves as a preset controlling parameter determining the energy differences between positive samples and negative samples. The final goal of QEB whose neural network architecture is shown in Figure 5 will be minimizing the total loss $L_{\text{total}}$ as follows,

$$L_{\text{total}} = L_{\text{head}} + L_{\text{tail}}.$$

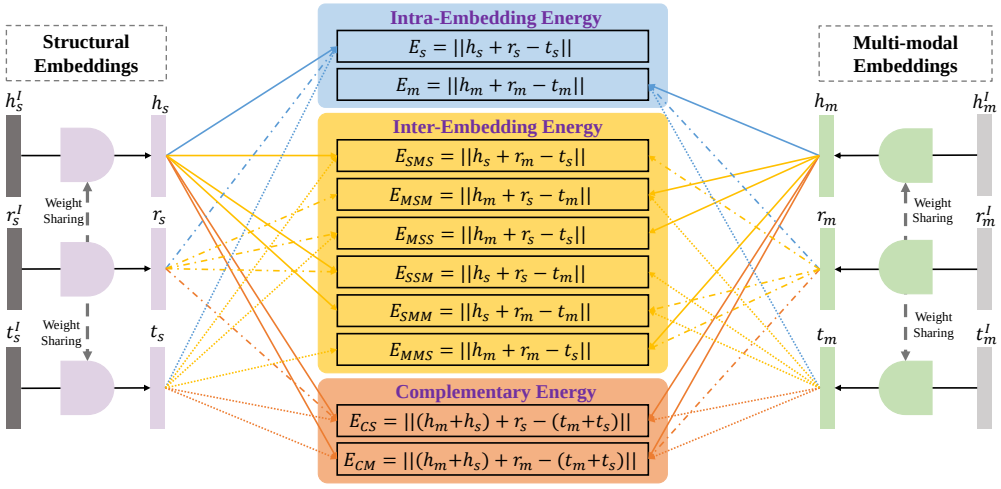


**Figure 5: Overview of the neural network architecture of our proposed QEB model.**

## 6 EXPERIMENTS

In this section, we conduct extensive experiments via comparing the performances of different state-of-the-art approaches as well as the proposed QEB model over our TIVA-KG, covering various scenarios ranging from unimodal to quadruple-modal settings. Necessary information for reproducing our results is available at https://github.com/Darkbblue/tiva-kg.

### 6.1 Experimental Settings

**Task.** We choose link prediction, the most widely adopted task for KG, to conduct experiments on. Same as other KGs, our TIVA-KG is composed of triplets (*head*, *relation*, *tail*) where the link prediction task aims to accurately predict *tail* given *head* and *relation* or predict *head* given *tail* and *relation*. During the training procedure, the ground truth *head* or *tail* normally will be replaced at random to generate negative samples.

**Datasets.** Following the common practice of existing works on large KGs, we extract a sub-graph from TIVA-KG as our datasets for experiments. The extracted sub-graph includes every entity and triplet within three hops from the entity "cat", containing 10K entities and 24K triplets. These triples are divided into a 20K training set, a 2K validation set and a 2K test set.

**Comparative Models.** We first examine the performances of four state-of-the-art models, TransE [3], TransD [11], DistMult [37] and NTN [30], to see if they can benefit from the quadruple modalities introduced by TIVA-KG. Given that these state-of-the-art unimodal approaches are designed to process only structural embeddings, we concatenate structural embeddings and multimodal embeddings together, transform them through a Multilayer Perceptron (MLP), and then feed the embeddings output via MLP into the models. We further examine the multimodal translation-based approach [21], with the same best hyperparameters reported in the work, as well as the same way of concatenating features from different modalities together. Finally, we examine our proposed QEB model, which is designed specifically for handling quadruple modalities in TIVA-KG.

**Evaluation Metrics.** We employ Hits@n and mean reciprocal rank (MRR) to evaluate the model performances for link prediction on TIVA-KG. Hits@n measures the ability to discover the ground truth result within top-n candidates, and MRR measures the average reciprocal of the rank of the ground truth in the predicted results. Larger Hits@n and MRR values indicate better performances.

**Multimodal Embeddings.** Different combinations of multimodal embeddings (i.e., text, text-image, text-image-video, text-image-video-audio) are employed to test the effects of utilizing multiple modalities. To combine multiple modalities together, we flatten and concatenate their features. They are concatenated in the order of text, image, video and audio, to provide the multimodal embeddings. If there are multiple instances for one modality, we simply use the first one and discard the others.

### 6.2 Experimental Results

We use "t, i, v, a" to denote text, image, video, audio, respectively, e.g., "tiv" means the tested model utilizes information from text, image and video modalities. "Unimodal" indicates that only the topological structure is taken into consideration.

**Unimodal Models.** The experimental results of four state-of-the-art unimodal models are shown in Table 5. We observe that taking multimodal knowledge into account can significantly improve model performances, because information from multiple modalities becomes available for utilization. However, these models cannot benefit from combining multiple modalities. For example, DistMult reaches the best performance at Hits@10 when considering all the four modalities, while achieves the best performances at Hit@1 and Hit@3 with only text and image modalities being taken into account. TransD and NTN even perform the best when only employing the text modality. Moreover, even the best results produced by TransE and DistMult are less than 50% at Hits@10, which shows that unimodal methods can not be naturally extended to multimodal scenarios, thus requiring further model designs to handle multimodal knowledge.

**Multimodal Models.** Table 6 and Table 7 demonstrate the experimental results of multimodal models, i.e., multimodal translation-based approach [21] and our proposed QEB model, in terms of

**Table 5: Results of unimodal models on link prediction. Results shown here are the average of head and tail predictions.**

| Model | Setting | Hits@1 | Hits@3 | Hits@10 | MRR |
|---|---|---|---|---|---|
| TransE | unimodal | 12.7 | 29.325 | 44.025 | 22.7 |
| | t | 43.65 | 46.575 | 47.9 | 45.4 |
| | ti | 38.775 | 43.25 | 45.4 | 41.4 |
| | tiv | 12.55 | 26.05 | 34.225 | 20.5 |
| | tiva | 0.0 | 0.225 | 0.925 | 0.41 |
| DistMult | unimodal | 27.575 | 34.35 | 34.95 | 31.0 |
| | t | 23.4 | 46.85 | 49.85 | 35.7 |
| | ti | **45.175** | **49.35** | 49.8 | **47.3** |
| | tiv | 44.15 | 49.225 | 49.9 | 46.7 |
| | tiva | 18.775 | 45.4 | **49.925** | 31.9 |
| TransD | unimodal | 4.45 | 24.825 | 40.7 | 16.5 |
| | t | 44.95 | 47.825 | 48.55 | 46.5 |
| | ti | 41.6 | 44.625 | 46.25 | 43.4 |
| | tiv | 19.45 | 45.0 | **49.925** | 32.2 |
| | tiva | 18.775 | 45.4 | **49.925** | 32.2 |
| NTN | unimodal | 30.75 | 44.82 | 47.325 | 37.7 |
| | t | 44.6 | 47.975 | 49.225 | 46.5 |
| | ti | 35.65 | 48.35 | 49.2 | 41.9 |
| | tiv | 37.15 | 48.35 | 49.375 | 42.8 |
| | tiva | 7.12 | 10.35 | 10.725 | 8.7 |

**Table 6: Results of multimodal models on (h,r,?) link prediction.**

| Model | Setting | Hits@1 | Hits@3 | Hits@10 | MRR |
|---|---|---|---|---|---|
| Multimodal Translation [21] | ti | 60.15 | 86.25 | 94.45 | 73.95 |
| | tiv | 59.3 | 85.4 | 95.5 | 73.04 |
| | tiva | 34.65 | 62.35 | 80.6 | 51.00 |
| | tiva-lstm | 66.9 | 89.6 | 96.65 | 78.79 |
| QEB (Ours) | ti | 80.6 | 92.55 | 97.1 | 87.00 |
| | tiv | **80.95** | 93.6 | 97.25 | **87.72** |
| | tiva | 62.8 | 79.8 | 91.45 | 72.93 |
| | tiva-lstm | 80.4 | **93.9** | **97.6** | 87.42 |

**Table 7: Results of multimodal models on (?,r,t) link prediction.**

| Model | Setting | Hits@1 | Hits@3 | Hits@10 | MRR |
|---|---|---|---|---|---|
| Multimodal Translation [21] | ti | 29.25 | 48.35 | 58.35 | 39.96 |
| | tiv | 40.8 | 65.2 | 74.25 | 54.32 |
| | tiva | 27.55 | 36.35 | 57.1 | 36.44 |
| | tiva-lstm | **59.3** | **75.45** | **84.15** | **68.26** |
| QEB (Ours) | ti | 45.1 | 59.65 | 73.5 | 55.05 |
| | tiv | 41.3 | 66.85 | 80.55 | 56.02 |
| | tiva | 10.0 | 16.65 | 42.4 | 19.01 |
| | tiva-lstm | 51.8 | 68.4 | 79.1 | 61.13 |

several settings for Hit@n and MRR. In addition to the "tiva" setting which utilizes audio features through simply padding or slicing them into representation with pre-defined length before flattening, we introduce an alternative setting "tiva-lstm" such that the audio modality can contribute to achieving better performances. In concrete, "tiva-lstm" processes audio embedding with an LSTM layer and an MLP layer, which provides better and more consistent results than "tiva" as shown in both Table 6 and Table 7.

It is obvious that both multimodal approaches perform much better than the unimodal methods shown in Table 5, which validates the capability of multimodal approaches in successfully capturing the interactions between different modalities to reach better performance. Empirical results under "ti", "tiv", "tiva-lstm" settings demonstrate a general trend of performance increase upon considering more modalities, which further proves the benefits of incorporating multimodal information on KG.

**Predict (h,r,?) v.s. Predict (?,r,t).** Through comparing the model performances in Table 6 and Table 7, we observe that predicting (?,r,t) is definitely more difficult. More importantly, the increase of "tiva-lstm" over "ti" for predicting (h,r,?), which is 0.48% of the original MRR. is less significant than that for predicting (?,r,t), which is 11.05% of the original MRR. This not only demonstrates the model performance boost brought by quadruple modalities, but also shows that incorporating information from multiple modalities may bring more benefits for more difficult tasks.

**QEB v.s. Multimodal Translation [21].** Multimodal Translation [21] can be regarded as a special case of our QEB model with only five energy terms (i.e., $E_s, E_{MSS}, E_{SSM}, E_{MSM}, E_{CS}$) and without the support to triplet grounding. On the one hand, Table 6 shows that QEB generally outperforms Multimodal Translation under all the four settings for (h,r,?) link prediction task. On the other hand, Table 7 implies that QEB performs better than Multimodal Translation under "ti" and "tiv" settings while worse under "tiva" and "tiva-lstm" settings for (?,r,t) link prediction task. The difference of performance gain on the two tasks indicates that different tasks on TIVA-KG may require contributions from different energy functions.

We conclude that quadruple modalities as well as triplet grounding can benefit link prediction task on Multimodal KGs, demonstrating that the two novel features of TIVA-KG can succeed in improving model performances for link prediction.

## 7 CONCLUSIONS AND FUTURE WORKS

We believe TIVA-KG has a great potential to promote the utilization of information from multiple modalities for knowledge mining on KGs. Although we propose QEB, a baseline model for TIVA-KG in this paper, what and how information from different modalities can be more elegantly combined to improve link prediction accuracy remain an interesting yet challenging problem. Furthermore, whether it is possible to employ TIVA-KG for other downstream tasks such as visual question answering, temporal sentence localization, multimedia search and recommendation to achieve performance boost also deserves future investigations.

## ACKNOWLEDGMENTS

This work was supported by the National Key Research and Development Program of China No. 2020AAA0106300, National Natural Science Foundation of China (No. 62222209, 62250008, 62102222), Beijing National Research Center for Information Science and Technology under Grant No. BNR2023RC01003, BNR2023TD03006, and Beijing Key Lab of Networked Multimedia.

# REFERENCES


[1] Houda Alberts, Ningyuan Huang, Yash Deshpande, Yibo Liu, Kyunghyun Cho, Clara Vania, and Iacer Calixto. 2021. VisualSem: a high-quality knowledge graph for vision and language. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*. 138–152.

[2] Ivana Balažević, Carl Allen, and Timothy Hospedales. 2019. TuckER: Tensor Factorization for Knowledge Graph Completion. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 5185–5194.

[3] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems* 26 (2013).

[4] Michel Buffa, Elena Cabrio, Michael Fell, Fabien Gandon, Alain Giboin, Romain Hennequin, Franck Michel, Johan Pauwels, Guillaume Pellerin, Maroua Tikat, et al. 2021. The WASABI dataset: Cultural, lyrics and audio analysis metadata about 2 million popular commercially released songs. In *European Semantic Web Conference*. Springer, 515–531.

[5] Takuma Ebisu and Ryutaro Ichise. 2018. Toruse: Knowledge graph embedding on a lie group. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[6] Sebastián Ferrada, Benjamín Bustos, and Aidan Hogan. 2017. IMGpedia: A Linked Dataset with Content-Based Analysis of Wikimedia Images. In *SEMWEB*.

[7] Luis Galárraga, Christina Teflioudi, Katja Hose, and Fabian M Suchanek. 2015. Fast rule mining in ontological knowledge bases with AMIE++. *The VLDB Journal* 24, 6 (2015), 707–730.

[8] Luis Antonio Galárraga, Christina Teflioudi, Katja Hose, and Fabian Suchanek. 2013. AMIE: association rule mining under incomplete evidence in ontological knowledge bases. In *Proceedings of the 22nd international conference on World Wide Web*. 413–422.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[10] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. CNN architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 131–135.

[11] Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Knowledge graph embedding via dynamic mapping matrix. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing*. 687–696.

[12] Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and S Yu Philip. 2021. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE transactions on neural networks and learning systems* 33, 2 (2021), 494–514.

[13] Seyed Mehran Kazemi and David Poole. 2018. Simple embedding for link prediction in knowledge graphs. *Advances in neural information processing systems* 31 (2018).

[14] Tamara G Kolda and Brett W Bader. 2009. Tensor decompositions and applications. *SIAM review* 51, 3 (2009), 455–500.

[15] Jonathan Lajus, Luis Galárraga, and Fabian Suchanek. 2020. Fast and exact rule mining with AMIE 3. In *The Semantic Web: 17th International Conference, ESWC 2020, Heraklion, Crete, Greece, May 31–June 4, 2020, Proceedings 17*. Springer, 36–52.

[16] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. 2020. Hierarchical conditional relation networks for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9972–9981.

[17] Guohao Li, Xin Wang, and Wenwu Zhu. 2020. Boosting visual question answering with context-aware knowledge aggregation. In *Proceedings of the 28th ACM International Conference on Multimedia*. 1227–1235.

[18] Hanxiao Liu, Yuexin Wu, and Yiming Yang. 2017. Analogical inference for multi-relational embeddings. In *International conference on machine learning*. PMLR, 2168–2178.

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

[20] Christian Meilicke, Melisachew Wudage Chekol, Daniel Ruffinelli, and Heiner Stuckenschmidt. 2019. Anytime Bottom-Up Rule Learning for Knowledge Graph Completion.. In *IJCAI*. 3137–3143.

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

[22] Dat Quoc Nguyen, Kairit Sirts, Lizhen Qu, and Mark Johnson. 2016. STransE: a novel embedding model of entities and relationships in knowledge bases. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 460–466.

[23] Maximilian Nickel, Lorenzo Rosasco, and Tomaso Poggio. 2016. Holographic embeddings of knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30.

[24] Daniel Oñoro-Rubio, Mathias Niepert, Alberto García-Durán, Roberto González-Sánchez, and Roberto J López-Sastre. 2018. Answering Visual-Relational Queries in Web-Extracted Knowledge Graphs. In *AKBC*.

[25] Meng Qu, Junkun Chen, Louis-Pascal Xhonneux, Yoshua Bengio, and Jian Tang. 2020. RNNLogic: Learning Logic Rules for Reasoning on Knowledge Graphs. In *International Conference on Learning Representations*.

[26] Luca Rossetto, Matthias Baumgartner, Narges Ashena, Florian Ruosch, Romana Pernisch, Lucien Heitz, and Abraham Bernstein. 2021. VideoGraph–towards using knowledge graphs for interactive video retrieval. In *MultiMedia Modeling: 27th International Conference, MMM 2021, Prague, Czech Republic, June 22–24, 2021, Proceedings, Part II*. Springer, 417–422.

[27] Andrea Rossi, Denilson Barbosa, Donatella Firmani, Antonio Matinata, and Paolo Merialdo. 2021. Knowledge graph embedding for link prediction: A comparative analysis. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 15, 2 (2021), 1–49.

[28] Ali Sadeghian, Mohammadreza Armandpour, Patrick Ding, and Daisy Zhe Wang. 2019. Drum: End-to-end differentiable rule mining on knowledge graphs. *Advances in Neural Information Processing Systems* 32 (2019).

[29] Ajeet Kumar Singh, Anand Mishra, Shashank Shekhar, and Anirban Chakraborty. 2019. From strings to things: Knowledge-enabled vqa model that can read and reason. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4602–4612.

[30] Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. 2013. Reasoning with neural tensor networks for knowledge base completion. *Advances in neural information processing systems* 26 (2013).

[31] Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 31.

[32] Robert Speer and Catherine Havasi. 2013. ConceptNet 5: A large semantic network for relational knowledge. *The People's Web Meets NLP: Collaboratively Constructed Language Resources* (2013), 161–176.

[33] Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2018. RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space. In *International Conference on Learning Representations*.

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

[35] Meng Wang, Guilin Qi, HaoFen Wang, and Qiushuo Zheng. 2019. Richpedia: A Comprehensive Multi-modal Knowledge Graph. In *Joint International Semantic Technology Conference*. Springer, 130–145.

[36] Ruobing Xie, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2017. Image-embodied knowledge representation learning. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. 3140–3146.

[37] Bishan Yang, Scott Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding Entities and Relations for Learning and Inference in Knowledge Bases. In *Proceedings of the International Conference on Learning Representations (ICLR) 2015*.

[38] Fan Yang, Zhilin Yang, and William W Cohen. 2017. Differentiable learning of logical rules for knowledge base reasoning. *Advances in neural information processing systems* 30 (2017).

[39] Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. 2022. Avqa: A dataset for audio-visual question answering on videos. In *Proceedings of the 30th ACM International Conference on Multimedia*. 3480–3491.

[40] Wen Zhang, Bibek Paudel, Wei Zhang, Abraham Bernstein, and Huajun Chen. 2019. Interaction embeddings for prediction and explanation in knowledge graphs. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 96–104.

[41] Xiangru Zhu, Zhixu Li, Xiaodan Wang, Xueyao Jiang, Penglei Sun, Xuwu Wang, Yanghua Xiao, and Nicholas Jing Yuan. 2022. Multi-modal knowledge graph construction and application: A survey. *IEEE Transactions on Knowledge and Data Engineering* (2022).