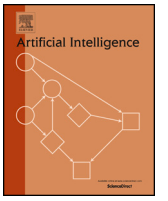
Artificial Intelligence 333 (2024) 104147

Contents lists available at ScienceDirect

Artificial Intelligence

journal homepage: www.elsevier.com/locate/artint



Knowledge is power: Open-world knowledge representation learning for knowledge-based visual reasoning ☆,☆☆

Wenbo Zheng [a,*], Lan Yan [b], Fei-Yue Wang [c]

[a] School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan 430070, China
[b] College of Computer Science and Engineering, Hunan University, Changsha 410082, China
[c] Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

ARTICLE INFO

Keywords:
Visual reasoning
Knowledge representation learning
Open-world learning
Graph model

ABSTRACT

Knowledge-based visual reasoning requires the ability to associate outside knowledge that is not present in a given image for cross-modal visual understanding. Two deficiencies of the existing approaches are that (1) they only employ or construct elementary and *explicit* but superficial knowledge graphs while lacking complex and *implicit* but indispensable cross-modal knowledge for visual reasoning, and (2) they also cannot reason new/*unseen* images or questions in open environments and are often violated in real-world applications. How to represent and leverage tacit multimodal knowledge for open-world visual reasoning scenarios has been less studied. In this paper, we propose a novel open-world knowledge representation learning method to not only construct implicit knowledge representations from the given images and their questions but also enable knowledge transfer from a *known* given scene to an *unknown* scene for answer prediction. Extensive experiments conducted on six benchmarks demonstrate the superiority of our approach over other state-of-the-art methods. We apply our approach to other visual reasoning tasks, and the experimental results show that our approach, with its good performance, can support related reasoning applications.

1. Introduction

Knowledge-based visual reasoning [1,4–6] demands a strong artificial intelligence (AI) model to not only parse a given image and its question but also understand related knowledge for correctly answering the question. For example, as shown in Fig. 1 (left), an AI model must recognize the many persons in the given image and the meaning of "this place" in the question while understanding the related knowledge between them; based on this knowledge, the model can find the precise answer (i.e., "shopping"). Despite the great success achieved by classic reasoning models [3], this task makes it more challenging for AI models to achieve the human-like ability of open-world cross-modal reasoning associated with implicit knowledge. As shown in Fig. 1 (second/fourth from the left), a classic model [3] may be aware that "this street" in another given image and its question has the same meeting as the street in the last sample and will directly adopt the answers of the past. Therefore, in such a cross-modal scenario, *how to build an AI-based*

☆ This article belongs to Special Issue: Open-World AI.
☆☆ Its early version [1] was presented at the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD), Virtual Event, Singapore, 2021.
* Corresponding author.
E-mail addresses: zwb2022@whut.edu.cn (W. Zheng), ylan@hnu.edu.cn (L. Yan), feiyue.wang@ia.ac.cn (F.-Y. Wang).

https://doi.org/10.1016/j.artint.2024.104147
Received 29 December 2022; Received in revised form 5 May 2024; Accepted 8 May 2024

Available online 13 May 2024

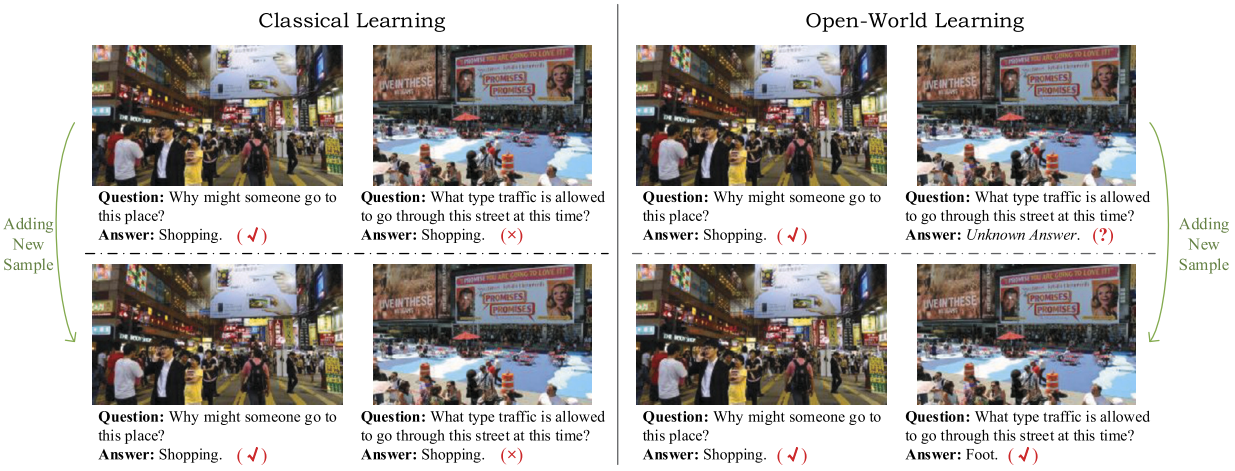


Fig. 1. An example of knowledge-based visual reasoning. The given samples (images and questions) are derived from the OK-VQA dataset [2]. The *left* answers are the outputs of our previous model [3]. The *right* answers are our new model's output.

reasoning model in the real world that can recognize things/knowledge it has seen/learned before, learn/master new things/knowledge it has not seen, and learn to adapt to new scenarios becomes a core problem of knowledge-based visual reasoning.

Nevertheless, most knowledge-based visual reasoning efforts [7–13] focus on the relations between visual and linguistic samples based on additional explicit knowledge, such as *structured* knowledge graphs (e.g., DBpedia [14], ConceptNet [15]) and *semistructured* knowledge bases (e.g., Wikipedia [16]). Without a doubt, these additional knowledge sources increase the information capacity of the constructed model. However, these knowledge sources usually simply represent objective facts that can only be described in natural language. Thus, it is not easy to describe complex scenarios with such explicit knowledge. For instance, an AI model must understand the behaviors of many persons appearing in the given image, as shown in Fig. 1 (first/third from the left), where classic learning-based models may grasp the relations between human, walk and foot. Nevertheless, these explicit knowledge sources seem to know what this place is, but lack such exemplary information, and may ignore the fact that "*this place*" is a walking street, which only allows walking.

On the other hand, some knowledge-based visual reasoning works [1,17–19] aim to construct a multimodal *structured* knowledge graph to model the knowledge between/across the visual and linguistic domains. However, this kind of approach still involves explicit knowledge in essence and cannot obtain complete/comprehensive (compositional complex and implicit/high-order logical) descriptions; they follow the closed-world assumption [20,21], where only *seen* samples appear during the reasoning process, and cannot work when adding new *unseen* samples.

In general, the current knowledge-based visual reasoning models have the following two limitations.

① The existing models are limited in terms of how they appropriately represent and leverage complex multimodal knowledge in the cross-modal knowledge-based visual reasoning scenario.

② It is challenging for such existing models to address given *known* samples while simultaneously being capable of realizing *unknown* samples and then continuing to update incrementally without retraining when adding new samples.

To this end, we propose a novel, effective and robust knowledge representation learning method, named **open-world knowledge representation learning** (OWKRL), for the knowledge-based visual reasoning task; this approach which exploits implicit visual knowledge throughout the whole process. In this work, on the one hand, we employ a comprehensive knowledge triplet (i.e., *head entity-relation-tail entity*) to represent complex multimodal knowledge and thereby overcome the first difficulty; in particular, we present a novel graph-based self-cross transformer to represent the given image and its question as well as the corresponding answer. Depending on this model, the representations of the image and question are embedded in the head entity, the implicit relation between the head and the tail is expressed by the special embedding derived from this model, and the factual answer is regarded as the tail entity. On the other hand, based on this knowledge triplet, we further propose a novel open-world learning paradigm to prevail over the second difficulty; concretely, the model first finds new *unknown* samples via clustering and then discovers samples more based on their similarity to the previous samples. This process follows the knowledge triplet representations from coarse to fine by contrasting positive and negative triplets. Experimental results obtained on six benchmarks demonstrate that our method achieves the best performance in comparison with other state-of-the-art approaches.

Our contributions can be briefly summarized as follows.

☼ We propose a novel open-world knowledge representation learning approach for the knowledge-based visual reasoning task. *To the best of our knowledge, this is the first attempt to combine open-world representation learning with knowledge triplets to explore knowledge-based visual reasoning.* Experimental results show that the proposed approach has strong robustness and outperforms similar methods.

☼ We propose a novel graph-based self-cross transformer model that performs deep multimodal fusion in the transformer architecture to produce a representation from the given samples for the task of knowledge-based visual reasoning. Experimental results show that this design method has strong robustness and outperforms similar methods.

☼ We offer an open-world knowledge representation learning strategy and an associated model that not only achieve state-of-the-art knowledge-based visual reasoning performance but are also effective on other reasoning tasks, e.g., science question answering and medical visual question answering.

2. Related work

In this section, we review the three aspects of this paper: knowledge-based visual reasoning, multimodal knowledge in visual reasoning, and open-world learning in computer vision.

Knowledge-based visual reasoning is challenging but vital for universal visual reasoning [4]. Most efforts [7–13] focus on "how to 'retrieve first and then read'" from structured knowledge graphs; they first retrieve related information from structured knowledge graphs and then perform explicit reasoning based on the obtained information. This kind of approach [7–13] integrates different explicit knowledge from various external knowledge resources, e.g., ConceptNet [15] and Wikipedia [16]. Notably, the knowledge resources upon which such methods rely are in essence purely natural language and simplistic knowledge graphs and cannot represent complex multimodal knowledge due to the missing information of other modalities. With the rise of large-scale pretrained models, some works [22–25] have treated large language models (e.g., GPT-3 [26]) as implicit knowledge resources and used them in the "retrieve first and then read" pipeline. However, this kind of knowledge resource still lacks information about other modalities. Aditya et al. [27] also pointed out that the lack of knowledge integration as well as higher-level reasoning capabilities in these methods still pose a hindrance. To this end, our previous work [1] took advantage of a related image to enhance the original knowledge graph. However, this task cannot be implemented for a complex scenario, especially when dealing with ineffable knowledge that cannot be described in language. More recently, we directly designed a knowledge triplet to represent the implicit knowledge in a multimodal scenario and further *proposed a graph-based visual reasoning transformer model* upon such explicit knowledge triplets.

Multimodal knowledge in visual reasoning is the key to the rapid development of both language-only and vision-language reasoning models [28–35], which have explored a diverse set of fusion strategies such as self-attention (e.g., [36,37]). Inspired by the success of the transformer model, most efforts [18,22–25,38,39] consider how to design a transformer-based framework between given images and their questions. These models are usually first pretrained on large-scale image-text datasets and then fine-tuned on *implicit* knowledge to complete related visual reasoning tasks. Some works [3,40] use hard-crafted interaction modules between given images and their questions for *explicit* reasoning. However, these works can *only* reason on *seen* samples and cannot *distinguish/reject unseen* samples without retraining the utilized models. To this end, *we propose a novel open-world learning method with our designed knowledge triplets* for *implicit* multimodal knowledge.

Open-world learning in computer vision has gained increasing attention. This type of learning in computer vision first starts with image classification [41]. Open-world image classification [42] aims at classifying images belonging to *known* classes during training while identifying examples of *unknown* classes. Furthermore, a multiclass classification approach [43] in the open world has been proposed. If the *unknown* classes remain unaddressed, the constructed model either misclassifies *seen* classes or classifies false classes. Recently, the open-world scenario has spread to other tasks, such as object detection [44,45] and semantic instance segmentation [46,47]. Inspired by the tremendous success of the above works, we propose a novel open-world learning method for knowledge-based visual reasoning. More importantly, *to the best of our knowledge, this is the first time that open-world learning has been utilized in knowledge-based visual reasoning.*

3. Open-world knowledge representation learning

In this section, we describe open-world knowledge representation learning. We first present the open-world setting, then propose a knowledge triplet representation acquisition process, and finally design a knowledge representation learning strategy.

3.1. Problem setup

This subsection presents the open-world learning setting and the definition of a knowledge representation.

The Open-World Learning Setting: To perform visual reasoning in the real world, the desired model must understand the given image and its question, recognize the answers it has learned before, grasp new knowledge that it has not seen, and learn to accommodate new scenarios. Compared with the classic learning technique that follows the closed-world setting [20,21], in this open-world learning setting, the answer to the question of interest may not have appeared during training.

On the other hand, the knowledge-based visual reasoning task [48,49] focuses on understanding and answering the given question using unstructured knowledge. Compared with most models [39,50,51] that follow the "retrieve first and then read" pipeline from structured explicit knowledge bases, we adopt knowledge representation learning for the constructed knowledge triple to capture implicit knowledge. From this view, we further consider visual reasoning as a knowledge graph reasoning problem incorporating the open-world learning setting, as follows.

Open-world knowledge representation reasoning requires a model $\mathcal{M}^{time}$ that aims to find a set of missing knowledge triplets $T^{time}_{test} = \{(\mathbf{h}^{time}, \mathbf{r}^{time}, \mathbf{?}) | \mathbf{h}^{time}, \mathbf{?} \in E^{time}_{test}, \mathbf{r}^{time} \in R\}$ during the testing phase when given a triplet set $T^{time}_{train} = \{(\mathbf{h}^{time}, \mathbf{r}^{time}, \mathbf{t}^{time}) | h^{time}, t^{time} \in E^{time}_{train}, r^{time} \in R\}$ during the training phase at time $time$, where E^{time}_{train} and E^{time}_{test} are the set and superset of entities, respectively; R is the set of relations; and $\mathbf{h}^{time}$, $\mathbf{r}^{time}$, and $\mathbf{t}^{time}$ represent the head entity, relation, and tail entity, respectively. Notably, the relationship between the two sets is $T_{train} \subseteq T_{test}$. For visual reasoning purposes, $\mathbf{h}^{time}$ denotes the features of the given image and its question, and $\mathbf{t}^{time}$ represents the answer to the corresponding question.

Fig. 2. The open-world learning setting. At time *time*, a reasoning model aims at finding knowledge triplets, i.e., $\mathbf{h}^{time}$ (the represented head entity) from the given images and their questions, $\mathbf{t}^{time}$ (the represented tail entity) when predicting the answers, and $\mathbf{r}^{time}$ (the represented relation). At this moment, the model in the testing stage fails to find a desired answer and must output "*unknown answer*". After performing incremental training, the model overwrites its previous answer as "foot" when adding new training samples.

In particular, the model $\mathcal{M}^{time}$ at time *time* is trained to recognize unlearned answers as belonging to the *unknown* vocabulary (denoted by the label "*unknown answer*"), in addition to answering the previously encountered *known* vocabularies from $\mathbf{t}^{time}$. These *unknown* answers recognized by $\mathcal{M}^{time}$ are forwarded to an oracle, which can label them and offer corresponding training samples. As a result, the provided training samples (images and questions) are incrementally added to $\mathbf{h}^{time}$, and $\mathbf{h}^{time+1}$ is obtained; similarly, the labeled answers are incrementally added to $\mathbf{t}^{time}$, and $\mathbf{t}^{time+1}$ is obtained. Then, $\mathcal{M}^{time}$ is incrementally trained, and an updated model $\mathcal{M}^{time+1}$ that can generate the desired answer from $\mathbf{t}^{time+1}$ is obtained. The above cycle continues during the life cycle of

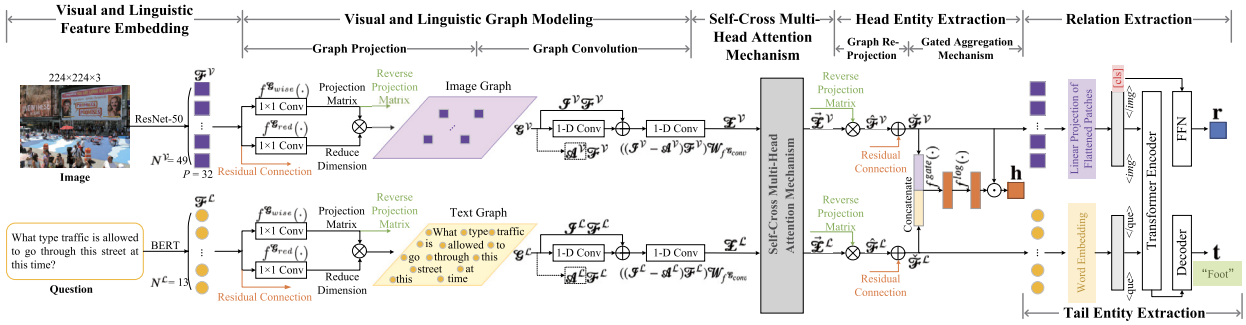


Fig. 3. An outline of the proposed knowledge triplet representation acquisition process, which includes two key steps: ① a novel graph-based self-cross transformer and ② a novel knowledge triplet extraction mechanism. For the former, we first use the pretrained models to obtain visual and linguistic features, which consist of the features $\mathfrak{F}^{\mathcal{L}}$ of the question and the image features $\mathfrak{F}^{\mathcal{V}}$; then, two graphs $\mathfrak{G}^{\mathcal{V}}$ and $\mathfrak{G}^{\mathcal{L}}$ are produced by projecting the visual and linguistic features; after that, these graphs perform graph convolution, and graph representations $\mathfrak{E}^{\mathcal{V}}$ and $\mathfrak{E}^{\mathcal{L}}$ are acquired. Finally, the proposed **self-cross multihead attention mechanism** shown in Fig. 4 is implemented; as a result, $\tilde{\mathfrak{E}}^{\mathcal{L}}$ and $\tilde{\mathfrak{E}}^{\mathcal{V}}$ are obtained. For the latter, we obtain $\tilde{\mathfrak{F}}^{\mathcal{V}}$ and $\tilde{\mathfrak{F}}^{\mathcal{L}}$ in the spatial domain by deriving them from the reprojection of these graph representations. Then, we introduce a gated aggregation mechanism to fuse the visual and linguistic features into a head representation; we offer an answer generator that produces an answer (e.g., "foot") as the tail representation; we use the special token of the answer generator as the relation representation.

the reasoning model, and the model can update itself and grasp new knowledge in each episode without forgetting the previously learned knowledge.

For example, as shown in Fig. 2, at time *time*, the model cannot generate an effective answer ("*unknown answer*") but can recognize that the target is an unlearned answer. Then, we add the training samples from another dataset [49] to form a new training set. After that, the model is incrementally trained, and the updated model can generate a correct answer ("foot").

3.2. Knowledge triplet representation acquisition

For knowledge-based visual reasoning, we treat the common sense knowledge in the complex scenario as implicit knowledge in the form of a knowledge triplet $(\mathbf{h}, \mathbf{r}, \mathbf{t})$ by following the setting mentioned in Section 3.1. The triplet construction process mainly consists of the following two steps, as shown in Fig. 3. 1) We design a novel graph-based self-cross transformer, aiming at the features of given images and their questions. 2) Based on this novel model, we extract a head entity, a relation, and a tail entity.

3.2.1. Graph-based self-cross transformer

In this subsection, we propose a graph-based self-cross transformer, as shown in Fig. 3. In particular, we first employ *visual and linguistic feature embedding* to obtain the features of a given image and its question and then design a *visual and linguistic graph modeling* method to build two semantic graphs. Furthermore, we offer a *self-cross multihead attention mechanism* to capture the relationships within a semantic graph and between semantic graphs.

❶ Visual and Linguistic Feature Embedding

In the first part, we take advantage of the pretraining method to extract the features of a given image and its question.

① **Visual Features**: An image with a size of $\mathcal{H} \times \mathcal{W} \times 3$ as an input can be divided into $N^{\mathcal{V}}$ patches ($N^{\mathcal{V}} = \frac{\mathcal{H} \times \mathcal{W}}{P^2}$, where P is the size of a patch), similar to the ViT [52]. Each patch is transformed into a feature vector $\boldsymbol{v}_i \in \mathbb{R}^{d_{N^{\mathcal{V}}}}$ (where $d_{N^{\mathcal{V}}}$ is the visual feature dimensionality) by the pretrained ResNet-50 model [53], and we obtain the visual features $\mathfrak{F}^{\mathcal{V}} = \{\boldsymbol{v}_i\}_{i=1}^{N^{\mathcal{V}}}$ as a result.

② **Linguistic Features**: Following our previous work [54], given an $N^{\mathcal{L}}$-word textual question, we employ the pretrained BERT model [55] to obtain linguistic features $\mathfrak{F}^{\mathcal{L}} = \{\boldsymbol{\ell}_i\}_{i=1}^{N^{\mathcal{L}}}$, where $\boldsymbol{\ell}_i \in \mathbb{R}^{d_{N^{\mathcal{L}}}}$ is the embedding of the i-th word, and $d_{N^{\mathcal{L}}}$ is the linguistic feature dimensionality.

❷ Visual and Linguistic Graph Modeling

In the second part, to capture the relationships among the words/image tokens in the given question and image, we construct visual/linguistic graphs via *graph projection*. Furthermore, we employ *graph convolution* to update the node representations acquired from these two graphs.

① **Graph Projection**: As shown in Fig. 3, we construct two graphs, an image graph $\mathfrak{G}^{\mathcal{V}}$ and a text graph $\mathfrak{G}^{\mathcal{L}}$, using the obtained features $\mathfrak{F}^{\mathcal{V}}$ and $\mathfrak{F}^{\mathcal{L}}$. For simplicity, we uniformly denote these two graphs as $\mathfrak{G}^{\mathcal{T}}$ and the original features as $\mathfrak{F}^{\mathcal{T}}$, where $\mathcal{T} \in \{\mathcal{V}, \mathcal{L}\}$. We project the features $\mathfrak{F}^{\mathcal{T}}$ into the graph $\mathfrak{G}^{\mathcal{T}} \in \mathbb{R}^{N^{\mathcal{T}} \times d_{N^{\mathcal{T}}}}$, where $N^{\mathcal{T}}$ is also the number of nodes and $d_{N^{\mathcal{T}}}$ is also the feature dimensionality of these nodes. As a result, the fully projected graph $\mathfrak{G}^{\mathcal{T}}$ is in essence a *lightweight fully connected* graph. To this end, we employ a linear transformation with its learnable weight, i.e., a projection function $f^{\mathfrak{G}_{proj}}(\cdot)$, as follows:

$$\begin{aligned} \mathfrak{G}^{\mathcal{T}} &= f^{\mathfrak{G}_{proj}}\left(f^{\mathfrak{G}_{red}}(\mathfrak{F}^{\mathcal{T}}; \mathfrak{W}_{f^{\mathfrak{G}_{red}}})\right) \\ &= f^{\mathfrak{G}_{wise}}(\mathfrak{F}^{\mathcal{T}}; \mathfrak{W}_{f^{\mathfrak{G}_{wise}}}) \times f^{\mathfrak{G}_{red}}(\mathfrak{F}^{\mathcal{T}}; \mathfrak{W}_{f^{\mathfrak{G}_{red}}}) \end{aligned} \tag{1}$$

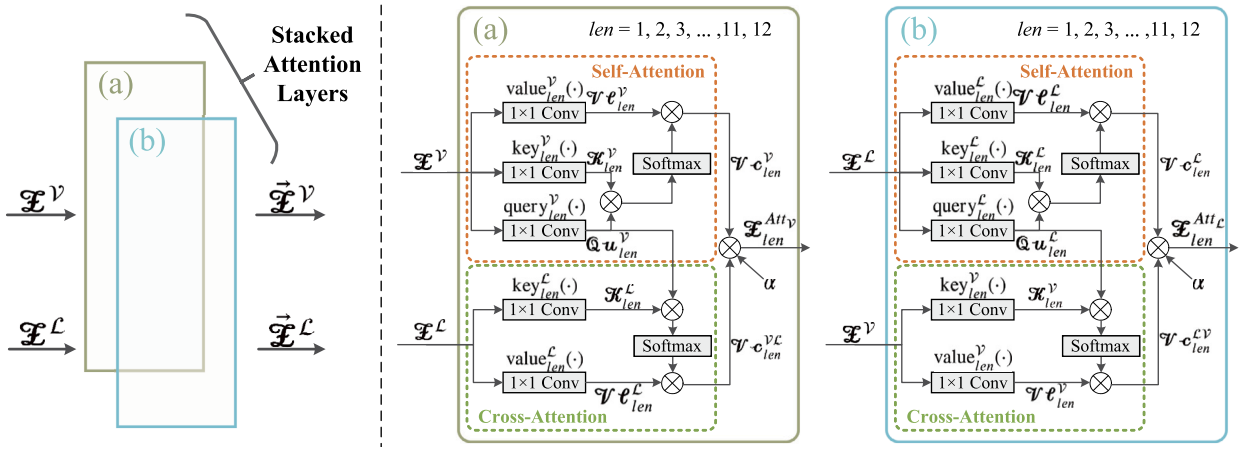


Fig. 4. A brief overview of self-cross multihead self-attention. Two kinds of attention layers are applied in a stacking fashion.

where we employ two convolution layers $f^{\mathcal{G}_{wise}}(\cdot)$ and $f^{\mathcal{G}_{red}}(\cdot)$ to achieve graph projection and feature dimensionality reduction, respectively, similar to the work of Ref. [56,57], and the weights of these two convolution layers are denoted as $\mathcal{W}_{f^{\mathcal{G}_{wise}}}$ and $\mathcal{W}_{f^{\mathcal{G}_{red}}}$, respectively.

② **Graph Convolution**: After completing the graph projection process, we obtain a projected graph $\mathcal{G}^{\mathcal{T}}$. Then, we employ the *graph convolution* [58] to propagate the information further and realize correlations between the relevant node features by learning the edge weights, as shown in Fig. 3. Specifically, a graph convolution with its parameters $\mathcal{W}_{f^{\mathcal{G}_{conv}}} \in \mathbb{R}^{d_{N^{\mathcal{T}}} \times d_{N^{\mathcal{T}}}}$ is specified as follows:

$$\begin{aligned} \mathcal{E}^{\mathcal{T}} &= f^{\mathcal{G}_{conv}}(\mathcal{A}^{\mathcal{T}} \mathcal{F}^{\mathcal{T}} \mathcal{W}_{f^{\mathcal{G}_{conv}}}) \\ &= ((\mathcal{I}^{\mathcal{T}} - \mathcal{A}^{\mathcal{T}})\mathcal{F}^{\mathcal{T}})\mathcal{W}_{f^{\mathcal{G}_{conv}}} \end{aligned} \tag{2}$$

where the adjacency matrix for cross-node diffusion in graph $\mathcal{G}^{\mathcal{T}}$ is denoted as $\mathcal{A}^{\mathcal{T}} \in \mathbb{R}^{N^{\mathcal{T}} \times N^{\mathcal{T}}}$ and the identity matrix is denoted as $\mathcal{I}^{\mathcal{T}} \in \mathbb{R}^{N^{\mathcal{T}} \times N^{\mathcal{T}}}$. To propagate the node features on the graph, a Laplacian smoothing operation [59] is applied. For the adjacency matrix $\mathcal{A}^{\mathcal{T}}$, we use gradient descent [60] to optimize its parameter $\mathcal{W}_{f^{\mathcal{G}_{conv}}}$. The identity matrix $\mathcal{I}^{\mathcal{T}}$ also contains a residual connection for each node in graph $\mathcal{G}^{\mathcal{T}}$.

❸ Self-Cross Multihead Attention Mechanism

Through the process of graph modeling, we obtain two graph embeddings from the given image and question: $\mathcal{E}^{\mathcal{V}}$ and $\mathcal{E}^{\mathcal{L}}$. We want to improve the single graph representation via its information while simultaneously capturing the relationship between two graph representations and using this relational representation to enhance the single graph representation. To this end, we design a cascaded multihead structure with stacked layers; it is an extension of the traditional multihead attention mechanism [55], as shown in Fig. 4, where each parallel head contains two kinds of attention: *self-attention* and *cross-attention*. As a result, we obtain final graph embeddings $\vec{\mathcal{E}}^{\mathcal{V}}$ and $\vec{\mathcal{E}}^{\mathcal{T}}$.

① **Self-Attention**: We replace the scaled dot product attention operation [61] in the original multihead attention mechanism [55] with a nonlocal attention block [62] as our self-attention layer, which focuses on improving the graph embedding via its own embedding. Specifically, for the len-th layer, we utilize two graph embeddings $\mathcal{E}^{t_1}_{len}$ and $\mathcal{E}^{t_2}_{len}$, where $t_1, t_2 \in \{\mathcal{V}, \mathcal{L}\}$, as the inputs of our self-attention mechanism, as follows:

$$\begin{aligned} \mathcal{Q}\mathcal{U}^{t_1}_{len} &= \text{query}^{t_1}_{len}(\mathcal{E}^{t_1}_{len}); \\ \mathcal{K}^{t_1}_{len} &= \text{key}^{t_1}_{len}(\mathcal{E}^{t_1}_{len}); \\ \mathcal{V}\mathcal{E}^{t_1}_{len} &= \text{value}^{t_1}_{len}(\mathcal{E}^{t_1}_{len}); \\ \mathcal{V}\mathcal{C}^{t_1}_{len} &= \text{softmax}((\mathcal{Q}\mathcal{U}^{t_1}_{len})^T \times \mathcal{K}^{t_1}_{len}) \times (\mathcal{V}\mathcal{E}^{t_1}_{len})^T, \end{aligned} \tag{3}$$

where we follow the classic scaled dot product self-attention process [61] and employ three linear transformations, $\text{query}^{t_1}_{len}(\cdot)$, $\text{key}^{t_1}_{len}(\cdot)$, and $\text{value}^{t_1}_{len}(\cdot)$, to obtain the attention vectors $\mathcal{Q}\mathcal{U}^{t_1}_{len}$, $\mathcal{K}^{t_1}_{len}$ and $\mathcal{V}\mathcal{E}^{t_1}_{len}$; the self-attention vector $\mathcal{V}\mathcal{C}^{t_1}_{len}$ is obtained by the softmax function [60].

② **Cross-Attention**: Similar to the above self-attention process, cross-attention aims at capturing the relationship between two graph embeddings and enhancing one graph embedding with the other graph embedding. To this end, we employ an asymmetric nonlocal neural block [63] to design our cross-attention mechanism, and this block can be represented as follows:

$$\begin{aligned}\mathcal{K}_{len}^{t_2} &= \text{key}_{len}^{t_2}(\mathcal{X}_{len}^{t_2});\\ \mathcal{V}\ell_{len}^{t_2} &= \text{value}_{len}^{t_2}(\mathcal{X}_{len}^{t_2});\\ \mathcal{V}\text{-}\mathcal{C}_{len}^{t_1 t_2} &= \text{softmax}((\mathcal{Q}\mathcal{U}_{len}^{t_1})^T \times \mathcal{K}_{len}^{t_2}) \times (\mathcal{V}\ell_{len}^{t_2})^T,\end{aligned} \tag{4}$$

where for the len-th layer, we make use of two linear transformations, $\text{key}_{len}^{t_2}(\cdot)$ and $\text{value}_{len}^{t_2}(\cdot)$, to obtain the attention vectors $\mathcal{K}_{len}^{t_2}$ and $\mathcal{V}\ell_{len}^{t_2}$, respectively; we utilize the softmax function to obtain the cross-attention vector $\mathcal{V}\text{-}\mathcal{C}_{len}^{t_1 t_2}$.

To associate self-attention with cross-attention, we include the gating mechanism shown in Fig. 3 to design the interactions between both kinds of attention:

$$\mathcal{X}_{len}^{Att_{t_1}} = \text{cat}((\mathcal{V}\text{-}\mathcal{C}_{len}^{t_1} + \alpha \times \mathcal{V}\text{-}\mathcal{C}_{len}^{t_1 t_2})^T, \mathcal{X}_{len}^{t_1}; \mathcal{W}_{\text{cat}}) \tag{5}$$

where we follow the research on nonlocal attention blocks in [62] and employ a 1×1 convolution as $\text{cat}(\cdot)$ with $\mathcal{W}_{\text{cat}}$ to adjust the nonlocal attention block and asymmetric nonlocal neural block; α is a learnable parameter initialized to 0.

In the last layer of the proposed self-cross multihead attention mechanism, we obtain two graph embeddings for the next step – graph reprojection, the size of which is consistent with the input of this proposed multihead attention mechanism – and we denote these embeddings as $\vec{\mathcal{X}}^{\mathcal{L}}$ and $\vec{\mathcal{X}}^{\mathcal{V}}$.

3.2.2. Head entity extraction

We define the head entity as the fusion of visual and linguistic feature embeddings that aggregate all input information. To this end, we first employ *graph reprojection* to obtain finer features and then design a *gated aggregation mechanism* to obtain the head entity representation **h**.

① **Graph Reprojection**: We use the projection matrix derived from the function $f^{\mathcal{G}_{wise}}(\cdot)$ mentioned in Section 3.2.1 (i.e., the projection matrix of $\mathcal{F}^{\mathcal{V}}$ and $\mathcal{F}^{\mathcal{L}}$). We follow the work of Ref. [56] and reproject the two graph embeddings $\vec{\mathcal{X}}^{\mathcal{L}}$ and $\vec{\mathcal{X}}^{\mathcal{V}}$ obtained through the proposed self-cross multihead attention mechanism into spatial features $\hat{\mathcal{F}}^{\mathcal{L}}$ and $\hat{\mathcal{F}}^{\mathcal{V}}$, respectively. Then, these spatial features are added with the residual connections of the features acquired from the inputs to obtain the final results $\check{\mathcal{F}}^{\mathcal{V}}$ and $\check{\mathcal{F}}^{\mathcal{L}}$.

② **Gated Aggregation Mechanism**: We follow our previous work [54] and take advantage of the gate mechanism to design an aggregation mechanism after the graph reprojection process. It aims to derive the visual feature from the given image that is most relevant to the corresponding question.

$$\mathbf{h} = f^{log}(f^{gate}(\check{\mathcal{F}}^{\mathcal{V}}, \check{\mathcal{F}}^{\mathcal{L}})) \odot \check{\mathcal{F}}^{\mathcal{V}} \tag{6}$$

where $\odot$ represents the elementwise multiplication operation, $f^{gate}(\cdot)$ denotes a two-layer stacked fully connected network that can concatenate two feature embeddings, and $f^{log}(\cdot)$ is the logistic sigmoid function.

3.2.3. Tail entity extraction

We define the tail entity as the answer generated according to the visual and linguistic features derived from the given image and its corresponding question. First, we employ the *graph reprojection* process mentioned in Section 3.2.2 to obtain finer features $\check{\mathcal{F}}^{\mathcal{V}}$ and $\check{\mathcal{F}}^{\mathcal{L}}$. Then, we utilize an answer generator by following the ViLT work [64], as shown in Fig. 3. The finer features $\check{\mathcal{F}}^{\mathcal{V}}$ and $\check{\mathcal{F}}^{\mathcal{L}}$ are fed into the answer generator to predict the output **t**. This task is regarded as a knowledge graph completion problem [65]. Formally, the answer generator $f^{AG}(\cdot)$ with its parameter $\mathcal{W}_{f^{AG}}$ is:

$$\mathbf{t} = f^{AG}(\check{\mathcal{F}}^{\mathcal{V}}, \check{\mathcal{F}}^{\mathcal{L}}; \mathcal{W}_{f^{AG}}) \tag{7}$$

① **Encoder**: We follow the ViLT work [64] and design an encoder based on the ViT [52] and BERT [55], as shown in Fig. 3. We employ $< img >$ and $< /img >$ to denote the start and end of the visual features $\check{\mathcal{F}}^{\mathcal{V}}$, respectively; similarly, $< que >$ and $< /que >$ are denoted as the start and end of the linguistic features $\check{\mathcal{F}}^{\mathcal{L}}$, respectively. We use a two-layer multilayer perceptron (MLP) head as a masked language model to output the answer **t** over the vocabulary.

② **Pretraining**: We pretrain the answer generator with 12 layers and a hidden size of 768. Following the ViLT work [64], we set the batch size to 4096 and pretrain for 100 K $\sim$ 200 K steps. We employ seven public datasets, GQA [66], Web QA [67], VQA 2.0 [68], Microsoft COCO (MSCOCO) [69], Visual Genome (VG) [70], SBU Captions (SBU) [71], and Google Conceptual Captions (GCC) [72], for pretraining, where the cross-entropy loss [60,73] is utilized.

3.2.4. Relation extraction

The relation is captured by the implicit relationship between the inputs (i.e., the given image and its question) and the output (i.e., the predicted answer). Thus, we can employ the answer generator (i.e., a hierarchical transformer) mentioned in Section 3.2.3 to achieve relation extraction. In particular, as shown in Fig. 3, we employ $[CLS]$ tokens and feed them into a two-layer feedforward network to obtain the relation representation **r**.

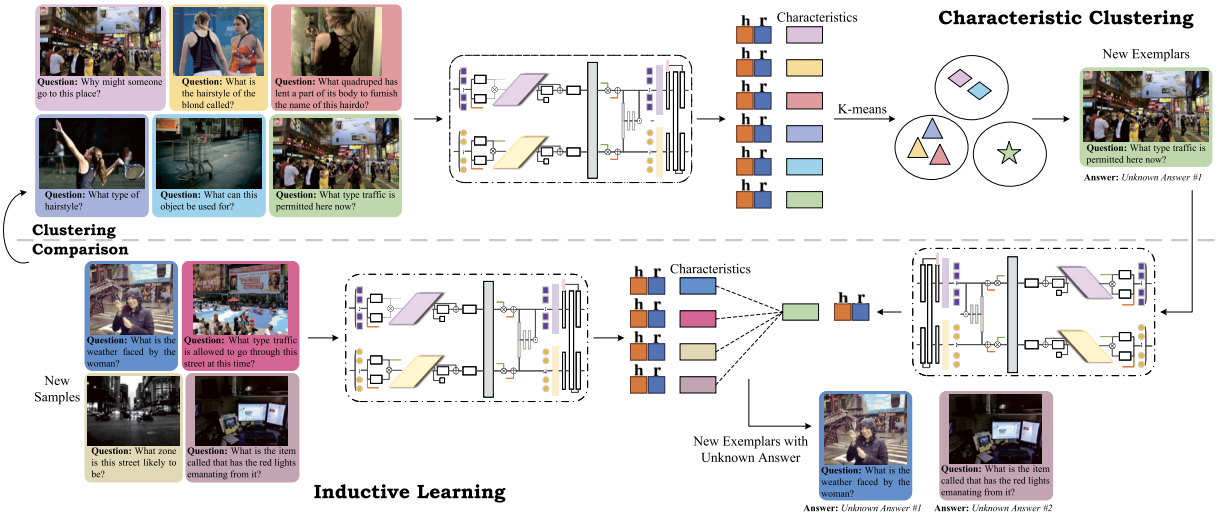


Fig. 5. A brief overview of the proposed open-world learning process, which consists of two stages: ① characteristic clustering and ② inductive learning. For the former, the model realizes *unknown* samples with the related examples by performing k-means clustering [74] on the characteristics derived from the head and relation representations. An element in a cluster represents a characteristic, while each color represents a given sample. The clusters in the ellipse belong to groups of categories, and each item in a cluster is an example. For the latter, the model discovers more examples by comparing the similarity between the new samples and previous examples. These two phases alternate to recognize and add unknown categories. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

3.3. Knowledge representation learning

In this subsection, we present the knowledge representation learning process of our model, as shown in Fig. 5. We introduce exemplar theory and design an exemplar-based learning strategy to achieve open-world learning. It aims to find textual answers with coherent characteristics belonging to the same *unknown* semantics. During this process, we further design related constraints to learn a knowledge triplet representation in a unified manner.

3.3.1. Open-world representation learning

Due to the open-world model learning setting mentioned in Sec. 3.1, we want to design a model that can train on the *known* samples/features and find and deal with *unseen/new* samples/features. To this end, we introduce exemplar theory [75,76] and propose a novel open-world representation learning technique for knowledge-based visual reasoning. In particular, we regard the *known* samples as exemplars and design a *characteristic clustering* strategy to find *unknown* samples associated with these exemplars via k-means clustering [74]. A closely bounded cluster with a high clustering score is considered a set of *unknown* samples, each of which is regarded as an exemplar. We compare each exemplar in this cluster with other exemplars in the subsequent minibatches according to their cosine similarities [77] to find new samples. These found samples also serve as pseudoground-truth answers (i.e., *unknown answer #*) for future iterations. In addition, we design an *inductive learning* mechanism to alleviate feature overfitting [78]. In general, as shown in Fig. 5, our open-world representation learning process consists of two steps.

① **Characteristic Clustering**: This step aims to find new samples and their exemplars in a minibatch. We use a four-layer stacked fully connected network to concatenate the head entity **h** and relation **r** and then obtain a 1024-dimensional fused characteristic for each sample in this minibatch. To find *unknown* samples, we perform k-means clustering [74] based on the cosine distance metric [79] using the extracted characteristics. In practice, we perform this approach every 200 iterations and employ the sample characteristics in the last 200 minibatches.

As a result, many clusters are generated, and we only take one subset of a relatively specific cluster corresponding to the *unknown* samples. Notably, such a cluster would have a high clustering score; however, it would have low mean cosine similarities between the centroid and all elements. This is because the characteristics of the *unknown* samples are easily clustered close together, while the *known* samples are loosely connected and have smaller scores. In practice, we generate 128 clusters in every 200 iterations for each minibatch. We select the top 10% clusters according to the average cosine similarity between the centroid and the cluster elements. We set the clustering score threshold to $0.92 \sim 0.99$. We find exemplars from these 10% clusters and memorize them in the subsequent stage.

② **Inductive Learning**: This process [78] has two purposes. On the one hand, this approach employs additional exemplars found in the past from the previous minibatches. To this end, we compare the characteristics of the exemplars in the memory with the characteristics in a new minibatch. On the other hand, to add more randomness, which can help to enhance the model's generalizability, we sample over the minibatches and only randomly select at most 20 samples from every minibatch for each model update. Overall, as shown in Fig. 5, these two stages alternate throughout the training procedure.

3.3.2. Knowledge triplet learning

Due to the knowledge triplet setting mentioned in Section 3.1, the head entity contains visual and linguistic information, the relation includes the relationship information between the image-question pair and its answer, and the tail entity has information about the answer. Notably, each element of a knowledge triplet represents a different sense, so we need to design a mechanism that can bridge the semantic and heterogeneous gaps. Inspired by image-embodied learning [80], we introduce a *structure-based representation constraint* to represent the closeness of the knowledge triplet for closing the heterogeneous gap. To construct the semantics between the head and tail entities, we further design a *task-specific representation constraint* to bridge the semantic gap.

① **Structure-Based Representation Constraint**: We use the TransE loss [81] to preserve the embedding structure by contrasting positive and negative triplets. Given image-question pairs, $\mathcal{Ans}^+$ and $\mathcal{Ans}^-$ are denoted as the correct (positive) and incorrect (negative) answer sets, respectively. A knowledge triplet $(\mathbf{h}, \mathbf{r}, \mathbf{t})$ represent extracted head entity (an image and its question), relation, and tail entity (answer) representations, respectively. We want to reduce the distance/divergence between each correct tail and $\mathbf{h}+\mathbf{r}$ and increase the distance between each incorrect tail and $\mathbf{h}+\mathbf{r}$.

$$Loss_{TransE} = \sum_{\mathbf{t}^+ \in \mathcal{Ans}^+} \sum_{\mathbf{t}^- \in \mathcal{Ans}^-} [\gamma + d(\mathbf{h}+\mathbf{r}, \mathbf{t}^+) - d(\mathbf{h}+\mathbf{r}, \mathbf{t}^-)]_+ \tag{8}$$

where $\mathbf{t}^+ \in \mathcal{Ans}^+$ denotes a positive answer, and $\mathbf{t}^- \in \mathcal{Ans}^-$ denotes a negative answer; $[\cdot]_+ \triangleq \max(0, \cdot)$; $d(\cdot, \cdot)$ represents the cosine distance following the work of Ref. [79]; γ is a learnable margin.

② **Task-Specific Representation Constraint**: To achieve knowledge-based visual reasoning, we employ the cross-entropy loss with negative sampling [82] by using a Monte Carlo approximation [83] as follows:

$$Loss_{CE} = \sum_{\mathbf{h},\mathbf{r},\mathbf{t}} (-\log(f^{\log}(-||\mathbf{h}+\mathbf{r}-\mathbf{t}||)) - \sum_{\mathbf{t}^- \in \mathcal{Ans}^-} [\log(f^{\log}(||\mathbf{h}+\mathbf{r}-\mathbf{t}^-||))]) \tag{9}$$

where $f^{log}(\cdot)$ is the logistic sigmoid function and $||\cdot||$ denotes the absolute value norm.

Overall, the total loss in this paper can be written as follows:

$$Loss = \mu \times Loss_{TransE} + \nu \times Loss_{CE} \tag{10}$$

where μ and ν are hyperparameters used to balance this loss, the best of which can be obtained by the NNI toolkit [84].

4. Experimental results and analyses

In this section, we conduct an experimental evaluation of the proposed approach on six benchmark datasets, and other state-of-the-art methods are compared with our approach in terms of performance. Note that "**Ours w/o OWL**" denotes a variant of our model and follows the classic training settings [1] without open-world learning.

4.1. Benchmark descriptions

In this subsection, we introduce the six benchmarks used for knowledge-based visual reasoning.

❶ The FVQA Dataset

The FVQA dataset [85] contains 2190 images and 5286 question-answer (QA) pairs and is further divided into training (2927 QA pairs) and testing (2899 QA pairs) sets.

❷ The KVQA Dataset

The KVQA dataset [86] contains 24602 images and their 183007 QA pairs, which are divided into two types according to the generated resources: original (ORG) and paraphrased (PRP) questions. We follow the training-validation-testing set splitting protocol of KVQA [86]: the dataset is randomly divided, with 70%, 20% and 10% of the images and their corresponding ∼ 130 K, ∼ 34 K and ∼ 19 K QA pairs used as the training, validation, and test sets, respectively.

❸ The Text-KVQA Dataset

The Text-KVQA dataset [9] contains 257 K images and their 1.3M QA pairs, and it is divided into three categories, i.e., scenes, movies, and books, according to the resources of the QA pairs. We follow the training-validation-testing set splitting protocol of Text-KVQA [9]: the dataset is randomly split, with 80%, 10%, and 10% of the images and their corresponding QA pairs used as the training, validation, and test sets, respectively.

❹ The OK-VQA Dataset

The OK-VQA dataset [2] contains 14031 images and 14055 QA pairs. We follow the training-validation-testing set splitting protocol of OK-VQA [2] and use the officially defined[1] training set, validation set, and test set.

❺ The A-OK-VQA Dataset

The A-OKVQA dataset [49] is an augmented successor of OK-VQA and contains 23.7 K unique images and 24903 QA pairs, which are split with 17.1K/1.1K/6.7 K images and their corresponding QA pairs for training, validation, and testing.

❻ The KRVQA Dataset

[1] https://okvqa.allenai.org/download.html.

Table 1
The comparison results obtained on the FVQA benchmark. The best model is marked in **bold**.

Model	Overall Accuracy (%)	
	Top-1	Top-3
LSTM-Question+Image+Pre-VQA [85]	24.98	40.40
BAN [91]	35.69	-
BAN +KG-Aug [7]	38.58	-
Hie-Question+Image+Pre-VQA [92]	43.14	59.44
SMCR [93]	48.90	-
UnifER+ViLT [24]	55.04	69.72
FVQA (top-3-Qqmaping) [85]	56.91	64.65
FVQA (Ensemble) [85]	58.76	—
Straight to the Facts (STTF) [94]	62.20	75.60
Reading Comprehension [85]	62.96	70.08
DPS [95]	63.56	76.47
KAN [96]	66.39	-
QAA [97]	68.74	-
Out of the Box (OB) [98]	69.35	80.25
Mucko [10]	73.06	85.94
PGVQA [99]	75.26	87.20
Hypergraph Transformer [22]	76.55	82.20
GRUC [100]	79.63	91.20
DMMGR+Dense Captioning [101]	81.20	95.38
Ours w/o OWL	83.45	97.34
Ours	**85.61**	**99.27**

The KRVQA dataset [87] contains 32910 images and 157201 QA pairs and is randomly divided into training, validation, and test sets with proportions of 60%, 20% and 20%, respectively, by following the official protocol [87]. In this dataset, the QA pairs are divided into two types (i.e., one-step and two-step pairs) and two other types (i.e., KB-related and KB-unrelated pairs) following the reasoning steps and the knowledge involvement strategy, respectively.

4.2. Experimental settings

In this subsection, we introduce the evaluation criteria for knowledge-based visual reasoning and describe the details of our implementation.

❶ Evaluation Criteria

We evaluate the performance achieved on the FVQA dataset using the top-1 and top-3 accuracy metrics. The average accuracy attained across 5 test splits is reported as the overall accuracy. Similar to the FVQA dataset, on the KVQA, Text-KVQA, and KRVQA datasets, we employ five such splits, and for the KVQA and KRVQA datasets, the average accuracy observed across five test splits is reported. For the Text-KVQA dataset, we follow our previous work [1] and use Recall@top-100 as the evaluation measure. For the OK-VQA and A-OK-VQA datasets, we use the top-1 accuracy.

The open-world learning setting is implemented as follows. Inspired by the evaluation settings of other open-world works/tasks [88,89], we statistically count all categories of answers in these six experimental datasets. We remove the annotations from the answer subset of the *known* categories in the training set and consider them *unknown answers*. We offer a general setting and construct the *unknown* splits with proportions of 5%.

❷ Implementation Details

We resize all input images to $224 \times 224 \times 3$. We use the AdamW optimizer [90]; the learning rate is initially set to 10^{-4}, and the weight decay is set to 10^{-2}. The patch size is set to 32, and patch projection yields 49 patches. The training batch size is set to 64. For our self-cross multihead attention, mechanism the number of heads is set to 8, and the number of stacked attention layers is set to 12.

4.3. Comparisons with the state-of-the-art approaches

In this subsection, we compare the state-of-the-art methods with our model on the six benchmarks. The compared methods, except for "Ours", follow the classic learning paradigm without the open-world learning setting.

❶ Comparison Conducted on the FVQA Dataset

On the FVQA benchmark, we compare our model with 19 state-of-the-art approaches, as described in Table 1.

Effect of Our Proposed Model. From Table 1, the performance of "Ours w/o OWL" is lower than that of Ours. This means that our open-world learning strategy is effective. On the other hand, the performance of "Ours w/o OWL" is better than that of the other approaches except for "Ours". This suggests that our model outperforms the other models, even without open-world learning. Thus, *our proposed model is effective* on the FVQA dataset for the knowledge-based visual reasoning task, both with and without the proposed open-world learning strategy.

Effect of Our Open-World Learning Strategy. From Table 1, the performance of "Ours" is the best. In particular, the performance of "Ours" is also better than that of "Ours w/o OWL". This reveals that our model is the most effective approach under

Table 2
The comparison results obtained on the KVQA benchmark. The best model is marked in **bold**.

Model	ORG	PRP	Mean
BLSTM [102]	48.0	27.2	37.6
MemNN [103]	50.2	34.2	42.2
GCN [58]	48.9	48.2	48.5
GGNN [104]	50.9	50.9	50.9
KVQAmeta [105]	-	-	52.83
HAN [106]	53.4	53.3	53.3
BAN [91]	59.6	60.0	59.8
Hypergraph Transformer [22]	62.0	62.8	62.4
Ours w/o OWL	64.5	65.1	64.8
Ours	**68.3**	**68.9**	**68.6**

Table 3
The comparison results obtained on the Text-KVQA Benchmark. The best model is marked in **bold**.

Method	Text-KVQA (scene)	Text-KVQA (book)	Text-KVQA (movie)
BoW + CNN [107]	11.5	8.7	7.0
BLSTM [102]	17.0	12.4	11.3
BLSTM + CNN [108]	19.8	17.3	15.7
HiCoAttenVQA [92]	22.2	20.2	18.4
BAN [91]	23.5	22.3	20.3
Memory network [8]	52.5	47.7	22.1
KEVQA [9]	54.5	49.8	23.0
HKEML [1]	60.2	71.2	52.3
Ours w/o OWL	67.3	78.5	62.4
Ours	**74.2**	**83.5**	**70.1**

the condition of open-world learning. Therefore, *our proposed open-world learning strategy is effective* on the FVQA dataset for the knowledge-based visual reasoning task.

❷ Comparison Conducted on the KVQA Dataset

On the KVQA benchmark, we compare our model with 8 state-of-the-art approaches, as described in Table 2.

Effect of Our Proposed Model. From Table 2, the performance of "Ours w/o OWL" is better than that of the other eight state-of-the-art methods in terms of ORG and PRP. This suggests that the proposed model is effective, even without open-world learning. Thus, *our proposed model is effective* on the KVQA dataset for the knowledge-based visual reasoning task, both with and without the proposed open-world learning strategy.

Effect of Our Open-World Learning Strategy. From Table 2, we also know that the performance of "Ours" is best among all state-of-the-art methods, and the performance of "Ours w/o OWL" is lower than that of "Ours". This shows that our proposed open-world learning strategy can truly yield improved knowledge-based visual reasoning performance. Therefore, *the proposed open-world learning strategy is effective* on the KVQA dataset for the knowledge-based visual reasoning task.

❸ Comparison Conducted on the Text-KVQA Dataset

On the Text-KVQA benchmark, we compare our model with 8 state-of-the-art approaches, including our previous work [1], as described in Table 3.

Effect of Our Proposed Model. From Table 3, the performance of "Ours w/o OWL" is better than that of the other eight approaches. In particular, the performance of "Ours w/o OWL" is better than that of HKEML (our previous method) [1]. This suggests that the proposed model is not only effective but also promotes and enhances "knowledge is power"-type models, even without open-world learning. Thus, *our proposed model is effective* on the Text-KVQA dataset for the knowledge-based visual reasoning task, both with and without the proposed open-world learning strategy.

Effect of Our Open-World Learning Strategy. From Table 3, it is obvious that "Ours" outperforms all state-of-the-art methods, and the performance of "Ours" is better than that of "Ours w/o OWL". This suggests that the proposed open-world learning strategy can truly yield improved knowledge-based visual reasoning performance. Therefore, *the proposed open-world learning strategy is effective* on the Text-KVQA dataset for the knowledge-based visual reasoning task.

❹ Comparison Conducted on the OK-VQA Dataset

On the OK-VQA benchmark, we compare our model with 38 state-of-the-art approaches, as described in Table 4.

Effect of Our Proposed Model. From Table 4, "Ours w/o OWL" outperforms the other state-of-the-art approaches, including our previous method (HKEML) [1] but excluding "Ours". This reveals that the proposed model is not only effective but can also extend and improve "knowledge is power"-type models, even when not using the open-world learning setting. Thus, *our proposed model is effective* on the OK-VQA dataset for the knowledge-based visual reasoning task, both with and without the proposed open-world learning strategy.

Effect of Our Open-World Learning Strategy. From Table 4, it is easily found that "Ours" has the best performance among all state-of-the-art methods and better performance than that of "Ours w/o OWL". This suggests that the proposed open-world learning strategy can truly yield improved knowledge-based visual reasoning performance. Therefore, *the proposed open-world learning strategy*

Table 4
The comparison results obtained on the OK-VQA benchmark. The best model is marked in **bold**.

Method	Top-1 Overall Accuracy (/%)	Method	Top-1 Overall Accuracy (/%)
Q-Only [48]	14.93	BAN/AN oracle [48]	27.59
MLP [48]	20.67	MUTAN/AN oracle [48]	28.47
BAN [91]	25.17	Mucko [10]	29.20
MUTAN [11]	26.41	KM^4 [3]	31.32
ArticleNet (AN) [48]	5.28	ViLBERT [38]	31.35
BAN+AN [48]	25.61	KRISP [109]	38.35
BAN + KG-Aug [7]	26.71	CONCAT [17]	31.95
GRUC [100]	29.87	ConceptBert [17]	33.66
MUTAN+AN [48]	27.84	HKEML [1]	39.24
LXMERT [39]	37.26	Unified-IO [23]	54.00
MAVEx [110]	39.40	SMCR [93]	30.46
PGVQA [99]	41.07	KAT [110]	54.41
Caption-DPR + Creader [111]	36.78	VLC-BERT [18]	43.14
Caption-DPR + Ereader [111]	39.20	PaLI [112]	54.5
KGE Aligning [12]	39.04	TRiG [13]	50.50
CBM + MMBERT [19]	39.20	MuKEA [65]	42.59
PICa [113]	43.30	UnifER [24]	42.13
T5+ Prefixes [39]	42.03	MetaLM [114]	46.5
LaKo [25]	47.01	Ours w/o OWL	58.9
PromptCap [115]	58.8	Ours	**62.1**

Table 5
The comparison results obtained on the A-OK-VQA benchmark. The best model is marked in **bold**.

Method	Overall Accuracy (/%)
Pythia [116]	21.9
ViLBERT [38]	25.9
LXMERT [39]	25.9
KRISP [109]	27.1
GPV-2 [117]	40.7
OFA Cap [118] + GPT-3 [26]	53.8
PromptCap [115]	58.0
Ours w/o OWL	59.8
Ours	**61.3**

is effective on the OK-VQA dataset for the knowledge-based visual reasoning task.

❺ Comparison Conducted on the A-OK-VQA Dataset

On the A-OK-VQA benchmark, we compare our model with 7 state-of-the-art approaches, as described in Table 5.

Effect of Our Proposed Model. From Table 5, the performance of "Ours w/o OWL" is better than that of the other approaches except for "Ours". This suggests that our model outperforms others, even without open-world learning. Thus, *our proposed model is effective* on the A-OK-VQA dataset for the knowledge-based visual reasoning task, both with and without the proposed open-world learning strategy.

Effect of Our Open-World Learning Strategy. From Table 5, the performance of "Ours" is the best. In particular, the performance of "Ours" is also better than that of "Ours w/o OWL". This reveals that our model is the most effective approach under the open-world learning condition. Therefore, *the proposed open-world learning strategy is effective* on the A-OK-VQA dataset for the knowledge-based visual reasoning task.

❻ Comparison Conducted on the KRVQA Dataset

On the KRVQA benchmark, we compare our model with 10 state-of-the-art approaches, as described in Table 6.

Effect of Our Proposed Model. From Table 6, "Ours w/o OWL" outperforms the other approaches except for "Ours". This suggests that our model is effective, even without open-world learning. Thus, *our model is effective* on the KRVQA dataset for the knowledge-based visual reasoning task, both with and without the proposed open-world learning strategy.

Effect of Our Open-World Learning Strategy. From Table 6, "Ours" has the best performance. The performance of "Ours" is better than that of "Ours w/o OWL". This suggests that our model is the most effective method in the open-world learning setting. Therefore, *the proposed open-world learning strategy is effective* on the KRVQA dataset for knowledge-based visual reasoning.

● Summary of the Comparison Results

From Table 1 ∼ Table 6, we find that our model is optimal and that the designed open-world learning strategy can boost the effectiveness of our model. Therefore, *our proposed model including open-world learning is effective* for the knowledge-based visual reasoning task.

4.4. Qualitative analysis

Since different competing approaches are available on each benchmark, we choose a prominent representative model for our qualitative analysis. In particular, we compare the proposed model with CLIP [124], BLIP-2 [125] and GPT-4 [126] on the six

Table 6
The comparison results obtained on the KRVQA benchmark. The best model is marked in **bold**.

Method	KB-unrelated							KB-related					Overall
	One-step			Two-step				One-step	Two-step				
	0	1	2	3	4	5	6	2	3	4	5	6	
Q-type [87]	36.19	2.78	8.21	33.18	35.97	3.66	8.06	0.09	0.00	0.18	0.06	0.33	8.12
LSTM [102]	45.98	2.79	2.75	43.26	40.67	2.62	1.72	0.43	0.00	0.52	1.65	0.74	8.81
FiLM [119]	52.42	21.35	18.50	45.23	42.36	21.32	15.44	6.27	5.48	4.37	4.41	7.19	16.89
MFH [120]	43.74	28.28	27.49	38.71	36.48	20.77	21.01	12.97	5.10	6.05	5.02	14.38	19.55
UpDn [121]	56.42	29.89	28.63	49.69	43.87	24.71	21.28	11.07	8.16	7.09	5.37	13.97	21.85
MCAN [122]	49.60	27.67	25.76	39.69	37.92	21.22	18.63	12.28	9.35	9.22	5.23	13.34	20.52
MuKEA [65]	59.12	44.88	37.36	52.47	48.08	35.63	31.61	17.62	6.14	9.85	6.22	18.28	27.38
Mucko [10]	-	-	-	-	-	-	-	-	-	-	-	-	24.00
KM-net [123]	-	-	-	-	-	-	-	-	-	-	-	-	25.19
DMMGR [101]	-	-	-	-	-	-	-	-	-	-	-	-	31.8
Ours w/o OWL	61.78	47.32	40.14	56.79	52.12	38.16	36.27	20.12	7.39	10.12	7.31	20.13	33.14
Ours	**64.38**	**51.37**	**44.24**	**60.14**	**57.35**	**42.15**	**41.27**	**22.34**	**9.82**	**12.35**	**9.24**	**22.34**	**36.42**

benchmarks. These methods, except for "Ours", follow the classic learning paradigm without the open-world learning setting. Besides, we use its payable API directly as GPT-4 is not open source by official way.[2]

The qualitative results are shown in Fig. 6. In order to better measure the difference in the outputs of all models, we statistically analyze the number of words in ground-truth for all the benchmarks and find that there are no more than 5 words, and therefore, we limit the outputs of all the models to no more than 5 words. For example, we input the textual question and its visual image as inputs directly into the GPT-4 API, with a unique instruction ("please provide an answer with no more than 5 words") and without other prompts for this kind of testing. From the sample in the first line from Fig. 6, it is clear that the output of GPT-4 is verbose and ambiguous, and even sometimes obviously false (last two from the right), compared with our model; compared with our model, the outputs of CLIP, BLIP-2, and "Ours w/o OWL" are both ambiguous. In the second line of the sample, it is evident that the GPT-4 cannot provide an answer at times, and when it can, it cannot do so precisely; CLIP and BLIP-2 give ambiguous or even false responses, sometimes with opposite random replies; while our model and its variant can offer the precise, accurate, and concise answers. CLIP and BLIP-2 from the third row of examples sometimes provide ambiguous and even wrong results compared to "Ours", "Ours w/o OWL" and GPT-4 that can provide accurate responses. In the other lines of the examples there are similar observations. From the above, the GPT-4, CLIP and BLIP-2 have been strong on other visual tasks, but unexpectedly, on the task of knowledge-based visual reasoning, a lack of open-world learning settings leads to the possibility of sometimes failing to response in a satisfying manner. Thus, *our proposed model including open-world learning is effective for the knowledge-based visual reasoning task*.

4.5. Performing reasoning with mismatched question

In this subsection, in order to further evaluate the knowledge-based reasoning ability of the model, we design this quality analysis experiment inspired by the work of Balepur et al. [127]. We purposefully break the combination of images and corresponding questions, where this break is to get a random combination except for correct pairings, and we directly input the images and corresponding disorderly questions into our model and the GPT-4 [126], which is considered to be a powerful large multi-modal model [128–131]. Similar to Sec. 4.4, we also limit the outputs of all the models to no more than 5 words.

The qualitative analysis results are shown in Fig. 7. The GPT-4 replies to each question, but these responses are sometimes incorrect, ambiguous, or even humorous (example in the top right corner). It makes one wonder whether the GPT-4 understands the given questions and its responses with reasoning, and may aim at ingratiating to humans with its answers. Compared to the GPT-4, our model mostly produces "unknown answer" for mismatched questions, thanking to the open-world learning settings. "To know what it is that you know, and to know what it is that you do not know– that is true understanding." [132] Obviously, our model's answers are more in line with the human comprehension of knowledge-based reasoning and understanding. Thus, *our proposed model including open-world learning is effective*.

4.6. Ablation study and model analyses

To verify the reasonableness and effectiveness of each part of our model, we design an ablation experiment on the OK-VQA and A-OK-VQA datasets. In Table 7, we report the results of our model and 11 variants. We analyze the following three aspects.

❶ Effect of Visual and Linguistic Graph Modeling

The graph modeling approach mentioned in Section 3.2.1 can transform the spatial features to features in a graph, and the goal is to capture the relationships within the features from visual/linguistic samples, which would enhance the representations of our model. From Table 7, it is clear that the methods with the image/text graph (#3, #4, #5, and #6) outperform those without the image/text graph (#1 and #2), regardless of whether open-world learning is used. This suggests that *the proposed visual and linguistic*

[2] We use the latest release gpt-4-turbo-preview (i.e., gpt-4-0125-preview) model.

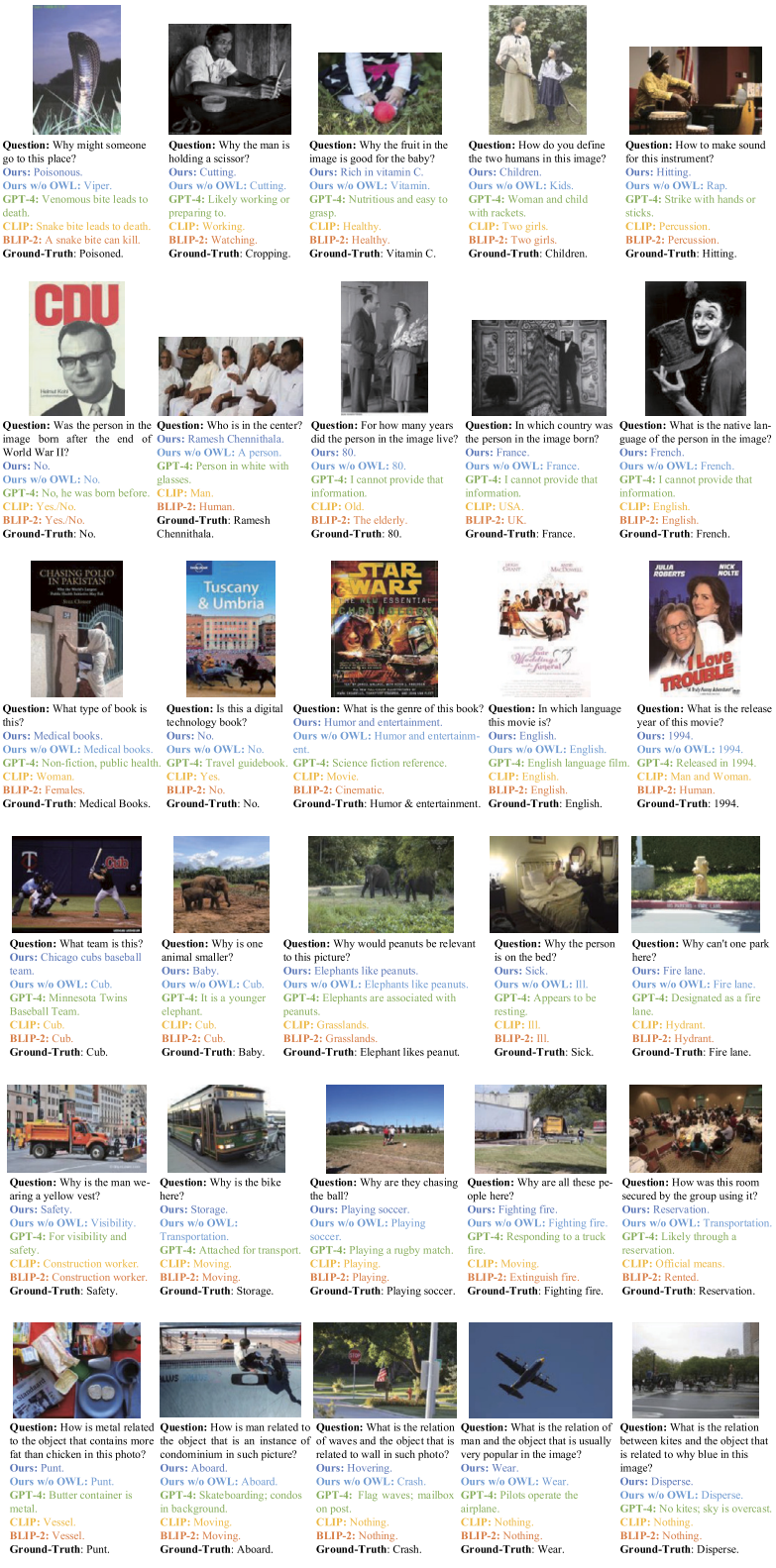
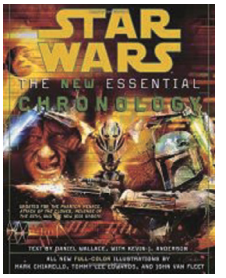


Fig. 6. Qualitative results of our model and CLIP [124], BLIP-2 [125] and GPT-4 [126]. The input samples for each row from top to bottom are from the FVQA, KVQA, Text-KVQA, OK-VQA, A-OKVQA and KRVQA benchmarks respectively.

Question: What is the relation of waves and the object that is related to wall in such photo?
Ours: Unknown Answer.
GPT-4: Snake scales resemble wave patterns.

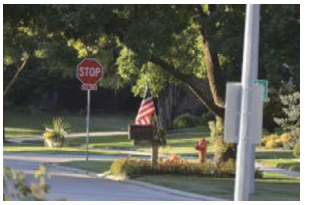
Question: Why are all these people here?
Ours: Party meeting.
GPT-4: Likely a formal political gathering.



Question: Why can't one park here?
Ours: Unknown Answer.
GPT-4: No parking in the Star Wars universe.

Question: What is the genre of this book?
Ours: Unknown Answer.
GPT-4: It's designated as a fire lane.

Question: Who is in the center?
Ours: Firefighter.
GPT-4: A person in a dark shirt is central.



Question: Why might someone go to this place?
Ours: Unknown Answer.
GPT-4: Residential visit or leisure walk.

Fig. 7. Reasoning results with mismatched question.

Table 7
The results of an ablation study conducted on the OK-VQA and A-OK-VQA datasets. √ represents a component that needs to be RESERVED; × represents a component that needs to be REMOVED. "w/o OWL" means that we do NOT use open-world learning, and "OWL" signifies the use of open-world learning.

#	Key Component					Top-1 Accuracy (/%)			
						OK-VQA Dataset		A-OK-VQA Dataset	
	Image Graph	Text Graph	Proposed Self-Cross Attention Mechanism	Self-Attention	Proposed Knowledge Triplet Learning Strategy	w/o OWL	OWL	w/o OWL	OWL
1	×	×	×	×	×	46.5	50.5	47.9	49.9
2	×	×	×	×	√	47.0	50.8	48.0	50.0
3	√	×	×	×	×	49.0	51.4	49.5	51.0
4	√	×	×	×	√	49.4	51.5	50.6	52.8
5	×	√	×	×	×	52.0	53.1	51.0	54.9
6	×	√	×	×	√	52.5	53.5	51.7	55.5
7	√	√	×	×	×	53.7	55.6	53.9	56.5
8	√	√	×	×	√	54.2	56.1	54.0	57.9
9	√	√	×	√	×	55.0	57.6	54.4	58.1
10	√	√	×	√	√	55.0	58.0	55.1	59.6
11	√	√	√	√	×	56.6	60.7	58.3	60.1
Ours	√	√	√	√	√	58.4	62.1	58.9	61.3

graph modeling approach is effective for the knowledge-based visual reasoning task.

❷ Effect of the Self-Cross Multihead Attention Mechanism

Our self-cross attention design mentioned in Section 2 has two kinds of attention modules: a self-attention module, which aims to improve the representations by itself, and a cross-attention module, which aims at incorporating one graph embedding mode into another graph embedding mode.

From Table 7, it is clear that the methods with self-cross attention (#11 and Ours) outperform those without self-cross attention (#7 and #8), regardless of whether open-world learning is considered. This suggests that *the whole designed self-cross attention mechanism is effective* for the knowledge-based visual reasoning task.

From Table 7, it is evident that the methods with self-attention (#9 and 10) outperform those without self-attention (#7 and #8), with or without open-world learning. This shows that *the self-attention mechanism is effective* for the knowledge-based visual reasoning task.

From Table 7, it is obvious that the methods with cross-attention (#11 and Ours) outperform those without cross-attention (#7 and #8), with or without open-world learning. This reveals that *the cross-attention mechanism is effective* for the knowledge-based visual reasoning task.

❸ Effect of Knowledge Triplet Learning

Our knowledge triplet learning approach mentioned in Section 3.3.2 is essential for capturing the tacit knowledge contained within the given datasets, and the goal is to achieve improved performance with this tacit knowledge. From Table 7, it is clear that the methods with knowledge triplet learning (#2, #4, #6, #8, #10, and Ours) outperform those without knowledge triplet learning

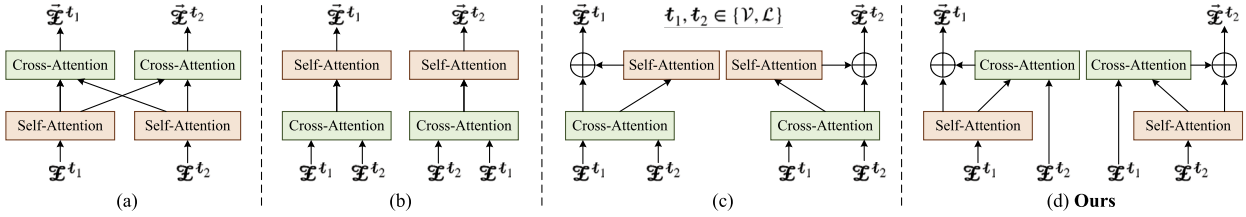
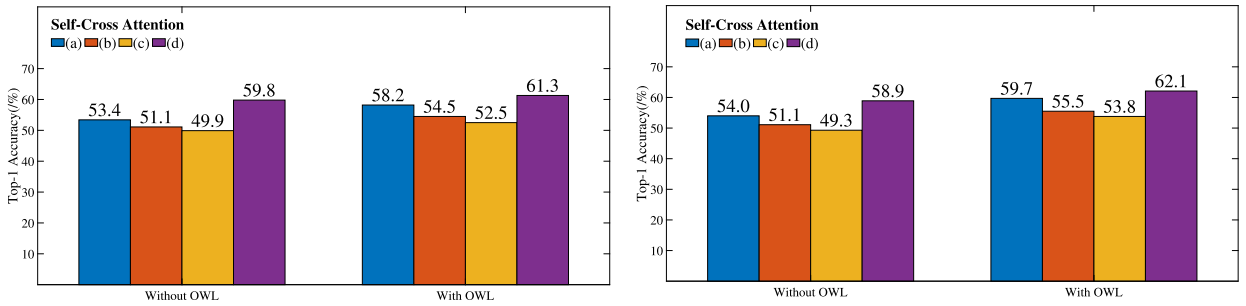


Fig. 8. Illustration of different self-cross attention mechanisms.



(a) OK-VQA Dataset; (b) A-OK-VQA Dataset.

Fig. 9. Comparison results obtained with different self-cross attention mechanisms.

(#1, #3, #5, #7, #9, and #11), with or without open-world learning. This suggests that *the designed knowledge triplet learning strategy is effective* for the knowledge-based visual reasoning task.

In general, this study illustrates a case of "knowledge is power" in the sense that the way that *knowledge triplets and the associated learning strategy are incorporated into our graph-based representation model* can be effective and improve the performance achieved in the knowledge-based visual reasoning task.

In addition, the methods with the proposed open-world learning strategy have better performance than those without open-world learning. This *also* suggests that *our open-world learning technique is effective* for the knowledge-based visual reasoning task.

4.7. Discussion of the self-cross attention mechanism

In this subsection, to validate the effectiveness of the proposed self-cross attention mechanism, we compare ours with different self-cross attention mechanisms on the OK-VQA and A-OK-VQA benchmarks. First, we describe these mechanisms in comparison with other self-cross attention mechanisms. Then, we analyze and discuss the empirical results.

In general, as shown in Fig. 8, different self-cross attention mechanisms that include the same self-attention and cross-attention modules have diverse combinations as follows.

(a): A graph embedding first passes through the self-attention mechanism, then passes through the cross-attention mechanism with another graph embedding, and finally obtains the corresponding result.

(b): As opposed to (a), two graph embeddings first pass through the cross-attention mechanism, then its result passes through the self-attention mechanism, and finally, two corresponding results are obtained.

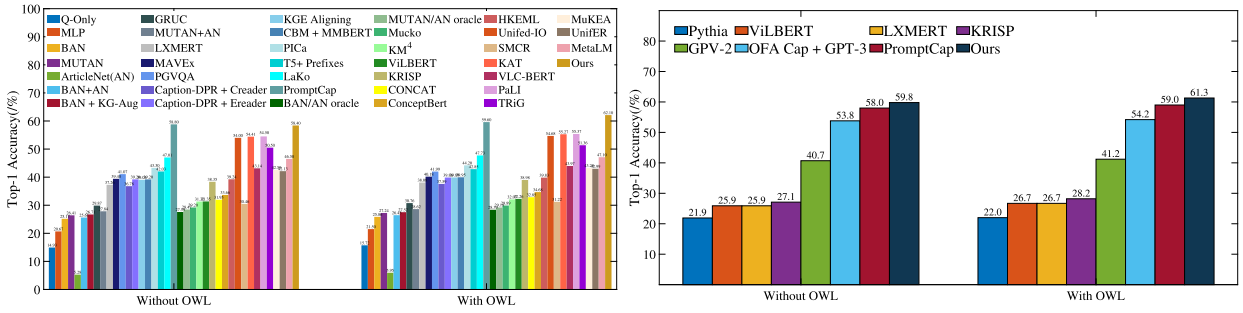
(c): Similar to (b), two graph embeddings first pass through the cross-attention mechanism, and its result passes through the self-attention mechanism. A gating mechanism weights the two above outcomes, and the corresponding result is derived.

(d): This is our proposed self-cross attention mechanism, which is detailed in Section 2. As opposed to (c), a graph embedding first passes through the self-attention mechanism and then through the cross-attention mechanism with another graph embedding. A gating mechanism weights the two above outcomes, and the corresponding result is finally derived.

For comparison purposes, we combine the above self-cross attention mechanisms with the proposed model to create several variants: our self-cross attention mechanism is replaced with (a), (b), and (c), and the remainder stays the same. Our variants are evaluated on the OK-VQA and A-OK-VQA datasets, and the empirical results are shown in Fig. 9.

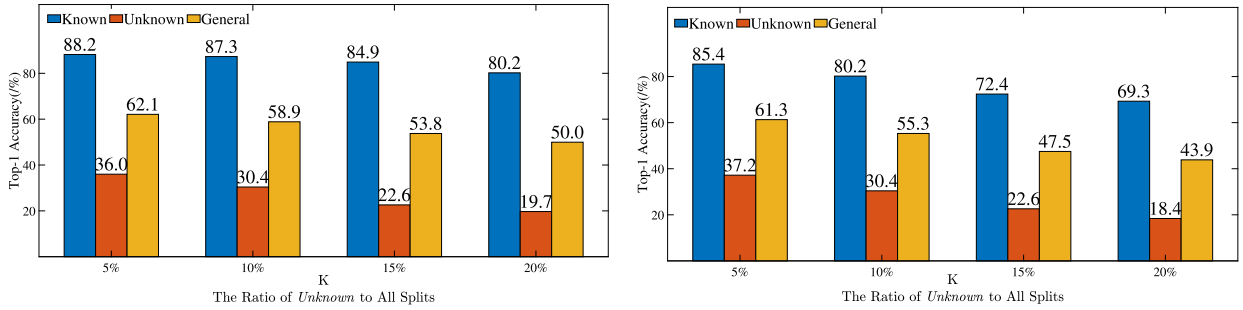
From Fig. 9, with or without open-world learning, the method with (d) has better performance those developed with (a), (b), and (c), and the methods with (d) and (a) have better performance than those developed with (c) and (b). Thus, the "self-attention first and then cross-attention" strategy is valid, which shows that *our proposed self-cross attention mechanism is effective* for the knowledge-based visual reasoning task.

In addition, the methods with the proposed open-world learning strategy have better performance than those without open-world learning. This *also* suggests that *our open-world learning approach is effective* for the knowledge-based visual reasoning task.



(a) OK-VQA Dataset; (b) A-OK-VQA Dataset.

Fig. 10. The generalization results.



(a) OK-VQA Dataset; (b) A-OK-VQA Dataset.

Fig. 11. Open-world knowledge reasoning results obtained with several different *known-unknown* splits, where K denotes the ratio of *unknown* to all samples.

4.8. Generalization of our open-world learning strategy

In this subsection, we design an experiment to further explore our proposed open-world learning method mentioned in Section 3.3.1. We follow the experimental settings of the comparison experiment conducted in Section 4.3, and we employ 39 state-of-the-art approaches on the OK-VQA dataset and 7 state-of-the-art approaches on the A-OK-VQA dataset, including our proposed model. We replace/improve the original training strategies of these methods and compel these methods to follow our open-world learning design. As a result, this kind of trained model is denoted as "With OWL"; we denote those trained in an authentic manner as "Without OWL".

These results are shown in Fig. 10. We observe different performance improvements for the various visual reasoning models in according to Fig. 10. This shows the generalizability of our proposed open-world learning strategy. Furthermore, this suggests that *our proposed open-world learning strategy is generalized and has a certain universality*.

4.9. Discussion of several different known-unknown splits

In this subsection, considering that the performance of open-world learning is closely related to the proportions of *unknown* splits, we design an experiment: we sort the answer categories based on their frequencies and regularly sample a subset of categories for removal to simulate *unknown* answers. Then, we construct four different splits with varying numbers of *unknown* answers (5%, 10%, 15%, and 20%). We separately report the performance of the model on the *known* and *unknown* sets, as well as the overall set (denoted as "general"), in Fig. 11.

From Fig. 11, it is clear that our model has the best performance when the ratio of unknown to all samples is set to 5. This suggests that *the open-world learning setting is available and effective* for the knowledge-based visual reasoning task.

In addition, from Fig. 11(a), our model still outperforms many state-of-the-art methods (33 approaches in Table 4) while using a large number (20%) of *unknown* samples; from Fig. 11(b), combined with Table 5, we can draw a similar conclusion. This *also* shows that *our proposed model and its open-world learning strategy are effective* for the knowledge-based visual reasoning task.

4.10. Utilizing the proposed model for the science question answering task

In this subsection, to better validate the robust performance of our model, we extend our model to the science question answering task. First, we describe the science question answering task and the utilized datasets. Later, we analyze and discuss the empirical results.

❶ Descriptions of the Task and Dataset

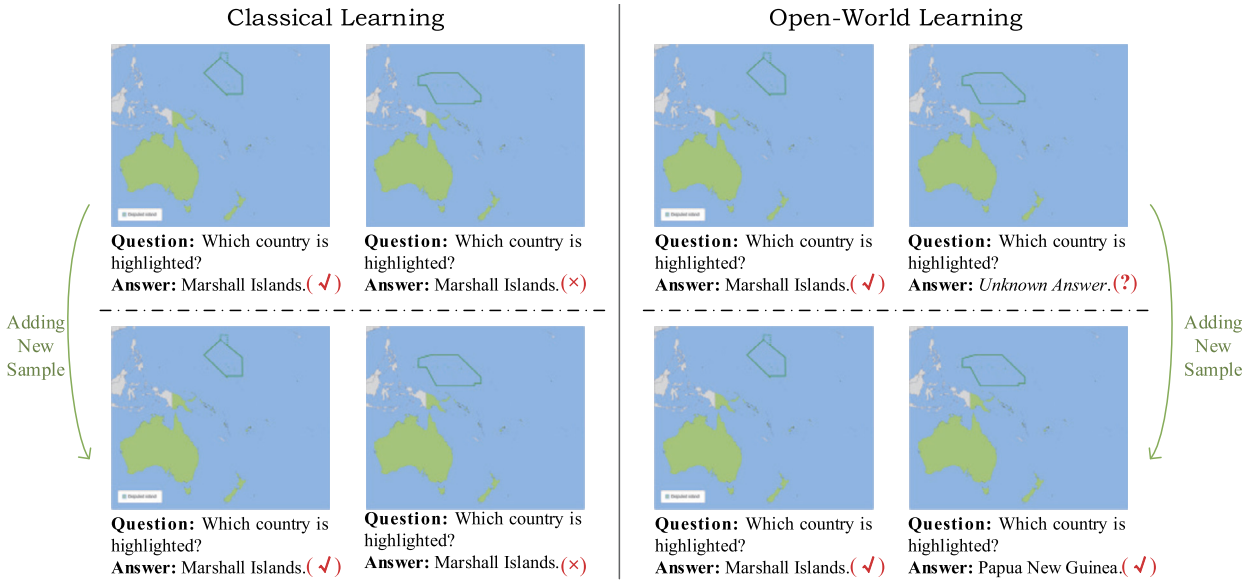


Fig. 12. The example of science question answering. The given samples (images and questions) are from the ScienceQA dataset [133]. The *left* answers are the outputs of our previous model [3]. The *right* answers are our current model's outputs.

Table 8
The comparison results obtained on the ScienceQA benchmark. Question classes: NAT = natural science, SOC = social science, LAN = language science, TXT = text context, IMG = image context, NO = no context, G1-6 = grades 1-6, G7-12 = grades 7-12. The best model is marked in **bold**.

Model	NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	Avg
Random chance	40.28	46.13	29.25	47.45	40.08	33.66	39.35	40.67	39.83
Q only [121]	41.34	27.22	47	41.79	35.15	44.6	39.28	40.87	39.85
MCAN [122]	56.08	46.23	58.09	59.43	51.17	55.4	51.65	59.72	54.54
Top-Down [121]	59.5	54.33	61.82	62.9	54.88	59.79	57.27	62.16	59.02
BAN [91]	60.88	46.57	66.64	62.61	52.6	65.51	56.83	63.94	59.37
DFAF [134]	64.03	48.82	63.55	65.88	54.49	64.11	57.12	67.17	60.72
ViLT [64]	60.48	63.89	60.27	63.2	61.38	57	60.72	61.9	61.14
Patch-TRM [135]	65.19	46.79	65.55	66.96	55.28	64.95	58.04	67.5	61.42
VisualBERT [136]	59.33	69.18	61.18	62.71	62.17	58.54	62.96	59.92	61.87
UnifiedQA [137]	68.16	69.18	74.91	63.78	61.38	77.84	72.98	65	70.12
GPT-3 [26]	75.44	70.87	78.09	74.68	67.43	79.93	78.23	69.68	75.17
Ours w/o OWL	76.12	71.58	79.23	76.12	70.12	82.34	80.45	72.34	76.04
Ours	**78.56**	**73.56**	**81.13**	**79.24**	**72.54**	**84.56**	**83.13**	**74.64**	**78.42**
Human	90.23	84.97	87.48	89.6	87.5	88.1	91.59	82.42	88.4

The science question answering task [133] requires that a model understand not only superficial multimodal information but also its implicit knowledge. The unique dataset for this task is ScienceQA [133], containing 21208 images and QA pairs derived from natural science, social science, and language science. We follow the training-validation-testing set splitting protocol of ScienceQA [133]: the dataset is split into training, validation, and test sets at a ratio of 60 : 20 : 20.

As shown in Fig. 12, for example, two images of nearly the same region are observed, and the same questions are questioned. Our proposed open-world learning strategy can answer correctly when adding new samples.

❷ Discussion of Science Question Answering

We adopt accuracy (%) as our evaluation metric. We conduct experiments with ten baselines to compare them with our proposed model.

Effect of Our Proposed Model. From Table 8, "Ours w/o OWL" outperforms the other approaches except for "Ours". This suggests that our model is effective, even without open-world learning. Therefore, *our proposed model is effective* for the science question answering task.

Effect of Our Open-World Learning Strategy. From Table 8, "Ours" has the best performance, and the performance of "Ours" is better than that of "Ours w/o OWL". This shows that "Ours" is the most effective approach in the open-world learning setting. Thus, *our proposed open-world learning strategy is effective* for the science question answering task.

From the two above points, we can conclude that our proposed model is effective, and our designed open-world learning strategy can help our model improve its science question answering performance. Our model has the best performance not only for knowledge-based visual reasoning but also for science question answering. Thus, *our model has robust performance*.

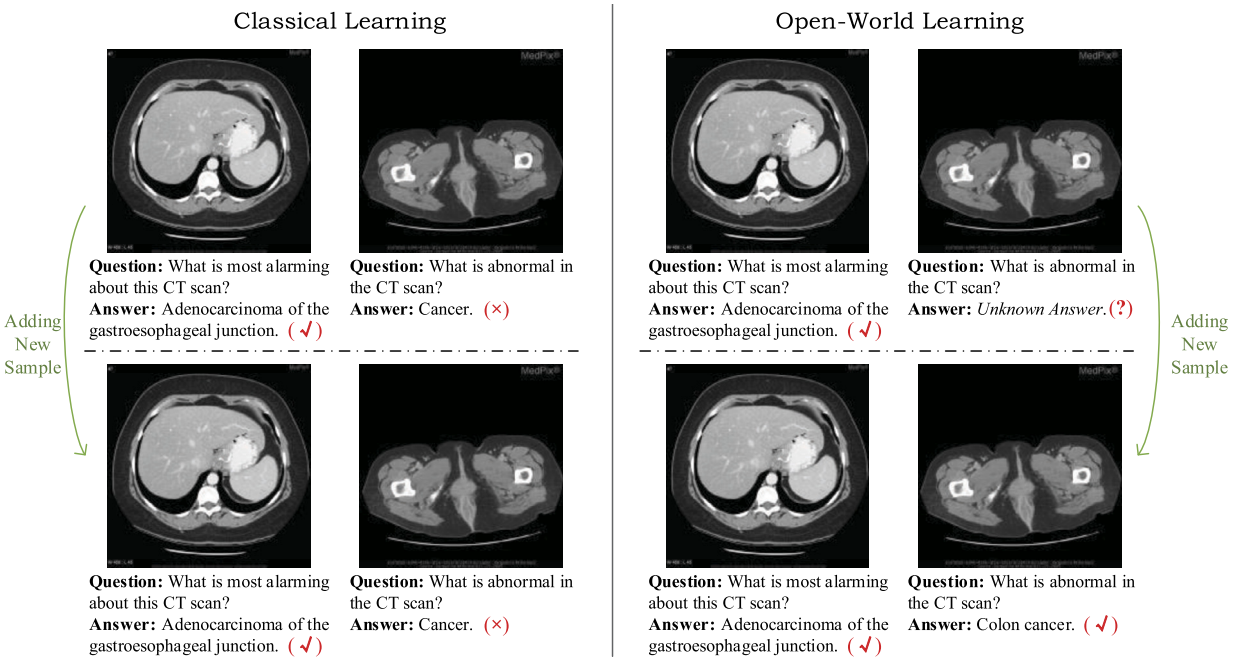


Fig. 13. Example of medical visual question answering. The given samples come from the Med-VQA-2021 dataset [138]. The *left* answers are the outputs of our previous model [3]. The *right* answers are the outputs of our current model.

Table 9
The comparison results obtained on the Med-VQA-2021 benchmark. The best model is marked in **bold**.

Method	Language Score (BLEU)	Classification Accuracy
SYSU-HCP [140]	0.416	0.382
Yunnan University [141]	0.402	0.362
TeamS [142]	0.391	0.348
Lijie [143]	0.352	0.316
IALab_PUC [144]	0.276	0.236
TAM [145]	0.255	0.222
Sheerin [146]	0.227	0.196
Ours w/o OWL	0.513	0.446
Ours	**0.546**	**0.489**

4.11. Utilizing the proposed model for the medical visual question answering task

In this subsection, to better validate the robust performance of our model, we extend our model to the medical visual question answering task. First, we describe the medical visual question answering task and the utilized dataset. Later, we analyze and discuss the empirical results.

❶ Descriptions of the Task and Dataset

The medical visual question answering task requires that a model understand a given medical image and grasp its related medical knowledge. We choose the Med-VQA-2021 dataset [138] for our experiment. In this dataset, the training set contains 4500 medical images and their 4500 QA pairs, the validation set consists of 500 medical images with 500 QA pairs, and the test set includes 500 medical images with 500 QA pairs.

As shown in Fig. 13, two radiological images of cancer are fed into the model, and their associated questions are expected to be answered. Our open-world learning strategy can answer correctly when adding new samples.

❷ Discussion of Medical Visual Question Answering

We adopt accuracy (%), and BLEU scores [139], which measure the similarity of the phrases, as our evaluation metrics. We conduct experiments with seven baselines to compare them with our proposed model.

Effect of Our Proposed Model. From Table 9, "Ours w/o OWL" outperforms the other seven approaches. This suggests that our proposed model is effective, even if it does not use open-world learning. Therefore, *our proposed model is effective* on the Med-VQA-2021 dataset for the medical visual question answering task.

Effect of Our Open-World Learning Strategy. From Table 9, "Ours" has the best performance among all methods, including "Ours w/o OWL". This shows that "Ours" with open-world learning is the most effective approach. Thus, *our open-world learning strategy is effective* on the Med-VQA-2021 dataset for the medical visual question answering task.

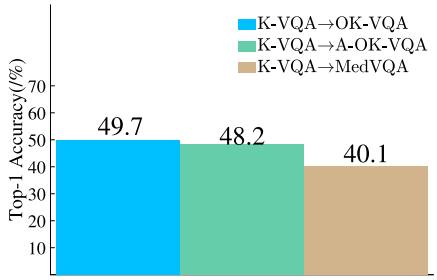


Fig. 14. Cross-dataset visual reasoning performance of the proposed approach, where A→B means that we pretrain our proposed approach on dataset A and that the model is trained in an end-to-end fashion on dataset B. The reported results are averaged over 10 independent runs.

From the two above points, we can conclude that the proposed model is valid, and we design an open-world learning method that can help the model improve its medical visual question answering performance. Our model has the best performance not only for knowledge-based visual reasoning but also for medical visual question answering. Thus, *our model has robust performance*.

4.12. Transfer learning ability of our open-world model

A challenge of our proposed model is to more smoothly combine learning and prior knowledge. If we give our model a problem it has not encountered before, we do not want the model to be powerless until it has been fed millions of labeled examples. The ideal model should be able to draw on what it already knows and use its transfer learning ability to apply that to the new problem. Accordingly, we devise transfer learning experiments. To explore the transfer learning ability of our model, we design two categories of experiments. First, we conduct an experiment on its transfer learning ability between content-related datasets. We pretrain our model on the K-VQA dataset [86] and then train and test our model on the OK-VQA [2] and A-OK-VQA [49] datasets separately. As stated in Sec. 4.1, we follow the training settings for our model pre-training/training on the K-VQA, OK-VQA and A-OK-VQA benchmarks, and also follow the testing settings for model test phase on the OK-VQA and A-OK-VQA benchmarks. Second, we perform experiments on its transfer learning ability between content-irrelevant datasets. To this end, we use the K-VQA dataset and the Med-VQA-2021 dataset [138]. Similarly, according to Sec. 4.1 and Sec. 4.11, we follow the training settings for our model pre-training/training on the K-VQA and Med-VQA-2021 datasets, and also adopt same testing settings as general model test on the Med-VQA-2021 dataset. Fig. 14 shows the results obtained with respect to the Acc. (%) evaluation metrics for ten independent runs on the datasets described earlier. *These experiments show that our method can achieve excellent visual reasoning performance in such a challenging scenario.*

☛ **Quantitative Analysis with Zero-Shot/Few-Shot Evaluation**

Further, we compare the efficacy of the methods pre-trained on K-VQA benchmark [86] with zero-shot and few-shot requirements on the OK-VQA [2], A-OK-VQA [49] and Med-VQA-2021 [138] datasets, where we utilize their respective official performance evaluation metrics, as stated in Sec. 4.1 and Sec. 4.11. The quantitative results are shown in Table 10.

We find that the zero/few-shot methods in Table 10 have a considerable performance margin compared to the approaches following the general training in Table 4, Table 5, and Table 9, due to the limited labeled samples and its significant content variations, but this is the kind of challenge we often face in our real-world daily without millions of labeled examples. Therefore, it is even more relevant to explore the transfer ability of the model's performance in this challenging scenario. From Table 10, in this complicated scenario, our model ("Ours") obtains the best performance under the zero-shot requirement by a significant margin ($\sim 5\%$ compared to hypergraph transformer [22]). Similarly, under the same few-shot requirement, our model ("Ours") obtains superior performance to other models. More importantly, it is noted that as the shot number increases, the performance of our model has a vast improvement. Besides, "Ours" is better than "Ours w/o OWL" on these three benchmarks. From the above, it is evident that *our proposed model including open-world learning is effective* for transfer learning on the task of knowledge-based visual reasoning.

5. Conclusion and future work

In this paper, we aim to solve two issues for the knowledge-based visual reasoning task: ① how to construct and exploit complex and implicit but indispensable multimodal knowledge in the cross-modal visual reasoning scenario and ② how to reject *unseen* samples that do not appear during training when incrementally learning these *unseen* samples to expand a model itself. To address these challenges, we propose a novel open-world knowledge representation learning strategy. In particular, on the one hand, we represent multimodal knowledge as a knowledge triplet to tie the visual and linguistic information derived from the given samples and the predicted/objective answers together with implicit relations; on the other hand, we present a novel open-world learning approach that can discriminate new samples by clustering and progressively discover unknown samples based on their similarity to the previous samples without retraining. Experimental results indicate that the proposed approach significantly outperforms other state-of-the-art approaches in the knowledge-based visual reasoning task on six benchmark datasets. In the future, on the one hand, we will advance the proposed approach by exploiting data from other fields (e.g., the Internet of Things or social network analysis); on the other hand, we will utilize richer information, especially some compelling human-designed features, during the graph interaction process.

Table 10
Transferring performance on OK-VQA, A-OKVQA and Med-VQA-2021 with the zero/few-shot requirements. The best model is marked in **bold**.

Methods	Shot Number	OK-VQA	A-OK-VQA	Med-VQA-2021	
		Top-1 Overall Accuracy	Overall Accuracy	Language Score (BLEU)	Classification Accuracy
Zero-Shot Evaluation					
BLSTM [102]	0	12.07%	10.95%	0.017	0.029
MemNN [103]	0	12.63%	11.21%	0.020	0.035
GCN [58]	0	12.90%	11.86%	0.024	0.039
GGNN [104]	0	13.48%	12.81%	0.028	0.043
KVQAmeta [105]	0	14.17%	13.61%	0.034	0.051
HAN [106]	0	15.30%	14.33%	0.042	0.063
BAN [91]	0	16.11%	15.11%	0.050	0.076
Hypergraph Transformer [22]	0	17.32%	17.24%	0.068	0.091
Ours w/o OWL	0	21.21%	20.34%	0.093	0.117
Ours	0	22.74%	22.15%	0.112	0.134
Few-Shot Evaluation					
BLSTM [102]	5	24.27%	23.10%	0.128	0.151
BLSTM [102]	10	26.86%	25.06%	0.153	0.171
MemNN [103]	5	24.74%	23.62%	0.134	0.155
MemNN [103]	10	27.12%	25.17%	0.157	0.172
GCN [58]	5	25.24%	24.04%	0.138	0.157
GCN [58]	10	27.93%	25.75%	0.165	0.176
GGNN [104]	5	26.71%	24.95%	0.147	0.167
GGNN [104]	10	27.99%	25.97%	0.168	0.178
KVQAmeta [105]	5	27.06%	25.13%	0.156	0.172
KVQAmeta [105]	10	29.08%	27.69%	0.190	0.193
HAN [106]	5	28.30%	26.26%	0.178	0.182
HAN [106]	10	30.44%	29.10%	0.209	0.205
BAN [91]	5	29.17%	27.88%	0.191	0.194
BAN [91]	10	31.55%	29.90%	0.214	0.212
Hypergraph Transformer [22]	5	30.09%	28.98%	0.207	0.204
Hypergraph Transformer [22]	10	32.52%	30.70%	0.222	0.219
Ours w/o OWL	5	32.21%	30.41%	0.217	0.217
Ours w/o OWL	10	36.17%	34.45%	0.266	0.251
Ours	5	34.46%	32.36%	0.244	0.237
Ours	10	38.99%	37.88%	0.298	0.284

Compliance with ethical standards

Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors.

CRediT authorship contribution statement

Wenbo Zheng: Writing – review & editing, Writing – original draft, Validation, Software, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization. **Lan Yan:** Writing – original draft, Methodology, Investigation. **Fei-Yue Wang:** Supervision, Resources, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This work is supported in part by the Natural Science Foundation of China (62303361, 62302161, and U1811463), in part by the Hainan Provincial Natural Science Foundation of China (623QN266), in part by the University-Industry Collaborative Education Program (231002531131826) and the Fundamental Research Funds for the Central Universities (WUT: 233110002), and in part by the National Key R&D Program of China (2018AAA0101502).

References

[1] W. Zheng, L. Yan, C. Gou, F.-Y. Wang, Knowledge is power: hierarchical-knowledge embedded meta-learning for visual reasoning in artistic domains, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, KDD '21, Association for Computing Machinery, New York, NY, USA, 2021, pp. 2360–2368.

[2] K. Marino, M. Rastegari, A. Farhadi, R. Mottaghi, Ok-vqa: a visual question answering benchmark requiring external knowledge, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[3] W. Zheng, L. Yan, C. Gou, F.-Y. Wang, KM^4: visual reasoning via knowledge embedding memory model with mutual modulation, Inf. Fusion (2021).

[4] Q. Wu, P. Wang, X. Wang, X. He, W. Zhu, Visual Question Answering-from Theory to Application, Springer, 2022.

[5] J. Suchan, M. Bhatt, S. Varadarajan, Commonsense visual sensemaking for autonomous driving – on generalised neurosymbolic online abduction integrating vision and semantics, Artif. Intell. 299 (2021) 103522, https://doi.org/10.1016/j.artint.2021.103522, https://www.sciencedirect.com/science/article/pii/S0004370221000734.

[6] İsmail İlkan Ceylan, A. Darwiche, G. Van den Broeck, Open-world probabilistic databases: semantics, algorithms, complexity, Artif. Intell. 295 (2021) 103474, https://doi.org/10.1016/j.artint.2021.103474, https://www.sciencedirect.com/science/article/pii/S0004370221000254.

[7] G. Li, X. Wang, W. Zhu, Boosting visual question answering with context-aware knowledge aggregation, in: Proceedings of the 28th ACM International Conference on Multimedia, MM '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 1227–1235.

[8] J. Weston, S. Chopra, A. Bordes, Memory networks, arXiv:1410.3916, 2014.

[9] A.K. Singh, A. Mishra, S. Shekhar, A. Chakraborty, From strings to things: knowledge-enabled vqa model that can read and reason, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

[10] Z. Zhu, J. Yu, Y. Wang, Y. Sun, Y. Hu, Q. Wu, Mucko: multi-layer cross-modal knowledge reasoning for fact-based visual question answering, in: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI'20, 2021.

[11] H. Ben-younes, R. Cadene, M. Cord, N. Thome, Mutan: multimodal tucker fusion for visual question answering, in: ICCV, 2017.

[12] V. Shevchenko, D. Teney, A. Dick, A. van den Hengel, Reasoning over vision and language: exploring the benefits of supplemental knowledge, in: Proceedings of the Third Workshop on Beyond Vision and LANguage: inTEgrating Real-World KNowledge (LANTERN), Association for Computational Linguistics, Kyiv, Ukraine, 2021, pp. 1–18, https://aclanthology.org/2021.lantern-1.1.

[13] F. Gao, Q. Ping, G. Thattai, A. Reganti, Y.N. Wu, P. Natarajan, Transform-retrieve-generate: natural language-centric outside-knowledge visual question answering, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5067–5077.

[14] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, Dbpedia: a nucleus for a web of open data, in: Proceedings of the 6th International the Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference, ISWC'07/ASWC'07, Springer-Verlag, Berlin, Heidelberg, 2007, pp. 722–735.

[15] R. Speer, J. Chin, C. Havasi, Conceptnet 5.5: an open multilingual graph of general knowledge, in: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI'17, AAAI Press, 2017, pp. 4444–4451.

[16] H. Singh, R. West, G. Colavizza, Wikipedia citations: a comprehensive data set of citations with identifiers extracted from English Wikipedia, Quant. Sci. Stud. 2 (1) (2021) 1–19, https://doi.org/10.1162/qss_a_00105, https://direct.mit.edu/qss/article-pdf/2/1/1/1906624/qss_a_00105.pdf.

[17] F. Gardères, M. Ziaeefard, B. Abeloos, F. Lecue, ConceptBert: concept-aware representation for visual question answering, in: Findings of EMNLP, 2020.

[18] S. Ravi, A. Chinchure, L. Sigal, R. Liao, V. Shwartz, VLC-BERT: visual question answering with contextualized commonsense knowledge, arXiv:2210.13626, 2022.

[19] A. Salaberria, G. Azkune, O. Lopez de Lacalle, A. Soroa, E. Agirre, Image captioning for effective use of language models in knowledge-based visual question answering, Expert Syst. Appl. 212 (2023) 118669, https://doi.org/10.1016/j.eswa.2022.118669, https://www.sciencedirect.com/science/article/pii/S0957417422017055.

[20] R. Reiter, On closed world data bases, in: B.L. Webber, N.J. Nilsson (Eds.), Readings in Artificial Intelligence, Morgan Kaufmann, 1981, pp. 119–140, https://www.sciencedirect.com/science/article/pii/B9780934613033500143.

[21] Z.-H. Zhou, Open-environment machine learning, Nat. Sci. Rev. 9 (8) (07 2022) nwac123, https://doi.org/10.1093/nsr/nwac123, https://academic.oup.com/nsr/article-pdf/9/8/nwac123/45957092/nwac123.pdf.

[22] Y.-J. Heo, E.-S. Kim, W.S. Choi, B.-T. Zhang, Hypergraph transformer: weakly-supervised multi-hop reasoning for knowledge-based visual question answering, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 373–390, https://aclanthology.org/2022.acl-long.29.

[23] J. Lu, C. Clark, R. Zellers, R. Mottaghi, A. Kembhavi, Unified-IO: a unified model for vision, language, and multi-modal tasks, arXiv:2206.08916, 2022.

[24] Y. Guo, L. Nie, Y. Wong, Y. Liu, Z. Cheng, M. Kankanhalli, A unified end-to-end retriever-reader framework for knowledge-based vqa, in: Proceedings of the 30th ACM International Conference on Multimedia, MM '22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 2061–2069.

[25] Z. Chen, Y. Huang, J. Chen, Y. Geng, Y. Fang, J. Pan, N. Zhang, W. Zhang, LaKo: knowledge-driven visual question answering via late knowledge-to-text injection, arXiv:2207.12888, 2022.

[26] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, vol. 33, Curran Associates, Inc., 2020, pp. 1877–1901, https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[27] S. Aditya, Y. Yang, C. Baral, Integrating knowledge and reasoning in image understanding, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, International Joint Conferences on Artificial Intelligence Organization, 2019, pp. 6252–6259.

[28] Z. Gan, L. Li, C. Li, L. Wang, Z. Liu, J. Gao, Vision-language pre-training: basics, recent advances, and future trends, Found. Trends® Comput. Graph. Vis. 14 (3–4) (2022) 163–352, https://doi.org/10.1561/0600000105.

[29] F. Chen, D. Zhang, M. Han, X. Chen, J. Shi, S. Xu, B. Xu, Vlp: a survey on vision-language pre-training, preprint, arXiv:2202.09061, 2022.

[30] Y. Du, Z. Liu, J. Li, W.X. Zhao, A survey of vision-language pre-trained models, preprint, arXiv:2202.10936, 2022.

[31] X. Zhu, Z. Li, X. Wang, X. Jiang, P. Sun, X. Wang, Y. Xiao, N.J. Yuan, Multi-modal knowledge graph construction and application: a survey, IEEE Trans. Knowl. Data Eng. (2022) 1–20, https://doi.org/10.1109/TKDE.2022.3224228.

[32] S. Khan, M. Naseer, M. Hayat, S.W. Zamir, F.S. Khan, M. Shah, Transformers in vision: a survey, ACM Comput. Surv. 54 (10s) (sep 2022), https://doi.org/10.1145/3505244.

[33] S. Uppal, S. Bhagat, D. Hazarika, N. Majumder, S. Poria, R. Zimmermann, A. Zadeh, Multimodal research in vision and language: a review of current and emerging trends, Inf. Fusion 77 (2022) 149–171, https://doi.org/10.1016/j.inffus.2021.07.009, https://www.sciencedirect.com/science/article/pii/S1566253521001512.

[34] A.A. Yusuf, F. Chong, M. Xianling, An analysis of graph convolutional networks and recent datasets for visual question answering, Artif. Intell. Rev. 55 (8) (2022) 6277–6300, https://doi.org/10.1007/s10462-022-10151-2.

[35] Y. Liu, Y.-S. Wei, H. Yan, G.-B. Li, L. Lin, Causal reasoning meets visual representation learning: a prospective study, Mach. Intell. Res. 19 (6) (2022) 485–511, https://doi.org/10.1007/s11633-022-1362-z.

[36] M.-H. Guo, T.-X. Xu, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, T.-J. Mu, S.-H. Zhang, R.R. Martin, M.-M. Cheng, S.-M. Hu, Attention mechanisms in computer vision: a survey, Comput. Vis. Media 8 (3) (2022) 331–368, https://doi.org/10.1007/s41095-022-0271-y.

[37] A. de Santana Correia, E.L. Colombini, Attention, please! A survey of neural attention models in deep learning, Artif. Intell. Rev. 55 (8) (2022) 6037–6124, https://doi.org/10.1007/s10462-022-10148-x.

[38] J. Lu, D. Batra, D. Parikh, S. Lee, Vilbert: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, vol. 32, Curran Associates, Inc., 2019, https://proceedings.neurips.cc/paper/2019/file/c74d97b01eae257e44aa9d5bade97baf-Paper.pdf.

[39] H. Tan, M. Bansal, LXMERT: learning cross-modality encoder representations from transformers, 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), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 5100–5111, https://aclanthology.org/D19-1514.
[40] S. Aditya, Y. Yang, C. Baral, Explicit reasoning over end-to-end neural architectures for visual question answering, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI'18/IAAI'18/EAAI'18, AAAI Press, 2018.
[41] J. Parmar, S.S. Chouhan, V. Raychoudhury, S.S. Rathore, Open-world machine learning: applications, challenges, and opportunities, ACM Comput. Surv. 55 (10) (2023) 1–37, https://doi.org/10.1145/3561381.
[42] W.J. Scheirer, A. de Rezende Rocha, A. Sapkota, T.E. Boult, Toward open set recognition, IEEE Trans. Pattern Anal. Mach. Intell. 35 (7) (2013) 1757–1772, https://doi.org/10.1109/TPAMI.2012.256.
[43] L.P. Jain, W.J. Scheirer, T.E. Boult, Multi-class open set recognition using probability of inclusion, in: D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), Computer Vision – ECCV 2014, Springer International Publishing, Cham, 2014, pp. 393–409.
[44] K.J. Joseph, S. Khan, F.S. Khan, V.N. Balasubramanian, Towards open world object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5830–5840.
[45] A. Gupta, S. Narayan, K.J. Joseph, S. Khan, F.S. Khan, M. Shah, Ow-detr: open-world detection transformer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9235–9244.
[46] J. Cen, P. Yun, J. Cai, M.Y. Wang, M. Liu, Deep metric learning for open world semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15333–15342.
[47] J. Xie, X. Hou, K. Ye, L. Shen, Clims: cross language image matching for weakly supervised semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4483–4492.
[48] K. Marino, M. Rastegari, A. Farhadi, R. Mottaghi, Ok-vqa: a visual question answering benchmark requiring external knowledge, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
[49] D. Schwenk, A. Khandelwal, C. Clark, K. Marino, R. Mottaghi, A-okvqa: a benchmark for visual question answering using world knowledge, in: S. Avidan, G. Brostow, M. Cissé, G.M. Farinella, T. Hassner (Eds.), Computer Vision – ECCV 2022, Springer Nature, Switzerland, Cham, 2022, pp. 146–162.
[50] J. Lu, D. Batra, D. Parikh, S. Lee, ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks, Curran Associates Inc., Red, Hook, NY, USA, 2019.
[51] K. Marino, X. Chen, D. Parikh, A. Gupta, M. 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 (CVPR), 2021, pp. 14111–14121.
[52] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: transformers for image recognition at scale, in: International Conference on Learning Representations, 2021, https://openreview.net/forum?id=YicbFdNTTy.
[53] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[54] W. Zheng, L. Yan, C. Gou, F.-Y. Wang, Two heads are better than one: hypergraph-enhanced graph reasoning for visual event ratiocination, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 139, PMLR, 2021, pp. 12747–12760, https://proceedings.mlr.press/v139/zheng21b.html.
[55] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186, https://aclanthology.org/N19-1423.
[56] Y. Chen, M. Rohrbach, Z. Yan, Y. Shuicheng, J. Feng, Y. Kalantidis, Graph-based global reasoning networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
[57] X. Liang, Z. Hu, H. Zhang, L. Lin, E.P. Xing, Symbolic graph reasoning meets convolutions, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems, vol. 31, Curran Associates, Inc., 2018, https://proceedings.neurips.cc/paper/2018/file/cbb6a3b884f4f88b3a8e3d44c636cbd8-Paper.pdf.
[58] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: International Conference on Learning Representations, 2017, https://openreview.net/forum?id=SJU4ayYgl.
[59] Q. Li, Z. Han, X.-m. Wu, Deeper insights into graph convolutional networks for semi-supervised learning, Proc. AAAI Conf. Artif. Intell. 32 (1) (Apr. 2018), https://doi.org/10.1609/aaai.v32i1.11604, https://ojs.aaai.org/index.php/AAAI/article/view/11604.
[60] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016, http://www.deeplearningbook.org.
[61] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L.u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, vol. 30, Curran Associates, Inc., 2017, https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
[62] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[63] Z. Zhu, M. Xu, S. Bai, T. Huang, X. Bai, Asymmetric non-local neural networks for semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019.
[64] W. Kim, B. Son, I. Kim, Vilt: vision-and-language transformer without convolution or region supervision, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 139, PMLR, 2021, pp. 5583–5594, http://proceedings.mlr.press/v139/kim21k.html.
[65] Y. Ding, J. Yu, B. Liu, Y. Hu, M. Cui, Q. Wu, Mukea: multimodal knowledge extraction and accumulation for knowledge-based visual question answering, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5089–5098.
[66] D.A. Hudson, C.D. Manning, Gqa: a new dataset for real-world visual reasoning and compositional question answering, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
[67] Y. Chang, M. Narang, H. Suzuki, G. Cao, J. Gao, Y. Bisk, Webqa: multihop and multimodal qa, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16495–16504.
[68] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, D. Parikh, Making the v in vqa matter: elevating the role of image understanding in visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[69] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: common objects in context, in: D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), Computer Vision – ECCV 2014, Springer International Publishing, Cham, 2014, pp. 740–755.
[70] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D.A. Shamma, M.S. Bernstein, L. Fei-Fei, Visual genome: connecting language and vision using crowdsourced dense image annotations, Int. J. Comput. Vis. 123 (1) (2017) 32–73, https://doi.org/10.1007/s11263-016-0981-7.
[71] V. Ordonez, G. Kulkarni, T. Berg, Im2text: describing images using 1 million captioned photographs, in: J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, K. Weinberger (Eds.), Advances in Neural Information Processing Systems, vol. 24, Curran Associates, Inc., 2011, https://proceedings.neurips.cc/paper/2011/file/5dd9db5e033da9c6fb5ba83c7a7ebea9-Paper.pdf.

[72] P. Sharma, N. Ding, S. Goodman, R. Soricut, Conceptual captions: a cleaned, hypernymed, image alt-text dataset for automatic image captioning, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 2556–2565, https://aclanthology.org/P18-1238.
[73] Z. Zhang, M. Sabuncu, Generalized cross entropy loss for training deep neural networks with noisy labels, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems, vol. 31, Curran Associates, Inc., 2018, https://proceedings.neurips.cc/paper/2018/file/f2925f97bc13ad2852a7a551802feea0-Paper.pdf.
[74] J. MacQueen, Classification and Analysis of Multivariate Observations, 1967.
[75] F.G. Ashby, L. Rosedahl, A neural interpretation of exemplar theory, Psychol. Rev. 124 (4) (2017) 472.
[76] J. Hwang, S.W. Oh, J.-Y. Lee, B. Han, Exemplar-based open-set panoptic segmentation network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1175–1184.
[77] Z.-H. Zhou, Machine Learning, Springer Nature, 2021.
[78] Q. Wu, C. Yang, J. Yan, Towards open-world feature extrapolation: an inductive graph learning approach, in: M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, J.W. Vaughan (Eds.), Advances in Neural Information Processing Systems, vol. 34, Curran Associates, Inc., 2021, pp. 19435–19447, https://proceedings.neurips.cc/paper/2021/file/a1c5aff9679455a233086e26b72b9a06-Paper.pdf.
[79] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, J. Gu, A strong baseline and batch normalization neck for deep person re-identification, IEEE Trans. Multimed. 22 (10) (2020) 2597–2609, https://doi.org/10.1109/TMM.2019.2958756.
[80] R. Xie, Z. Liu, H. Luan, M. Sun, Image-embodied knowledge representation learning, in: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17, 2017, pp. 3140–3146.
[81] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, O. Yakhnenko, Translating embeddings for modeling multi-relational data, in: C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K. Weinberger (Eds.), Advances in Neural Information Processing Systems, vol. 26, Curran Associates, Inc., 2013, https://proceedings.neurips.cc/paper/2013/file/1cecc7a77928ca8133fa24680a88d2f9-Paper.pdf.
[82] H. Kamigaito, K. Hayashi, Unified interpretation of softmax cross-entropy and negative sampling: with case study for knowledge graph embedding, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, 2021, pp. 5517–5531, Online, https://aclanthology.org/2021.acl-long.429.
[83] W.L. Hamilton, Graph representation learning, Synth. Lect. Artif. Intell. Mach. Learn. 14 (3) (2020) 1–159.
[84] Microsoft, Microsoft/NNI: an open source automl toolkit for automate machine learning lifecycle, including feature engineering, neural architecture search, model compression and hyper-parameter tuning, GitHub, https://github.com/microsoft/nni.
[85] P. Wang, Q. Wu, C. Shen, A. Dick, A. van den Hengel, Fvqa: fact-based visual question answering, IEEE Trans. Pattern Anal. Mach. Intell. 40 (10) (2018) 2413–2427, https://doi.org/10.1109/TPAMI.2017.2754246.
[86] S. Shah, A. Mishra, N. Yadati, P.P. Talukdar, Kvqa: knowledge-aware visual question answering, Proc. AAAI Conf. Artif. Intell. 33 (01) (2019) 8876–8884, https://doi.org/10.1609/aaai.v33i01.33018876, https://ojs.aaai.org/index.php/AAAI/article/view/4915.
[87] Q. Cao, B. Li, X. Liang, K. Wang, L. Lin, Knowledge-routed visual question reasoning: challenges for deep representation embedding, IEEE Trans. Neural Netw. Learn. Syst. 33 (7) (2022) 2758–2767, https://doi.org/10.1109/TNNLS.2020.3045034.
[88] A. Gupta, S. Narayan, K.J. Joseph, S. Khan, F.S. Khan, M. Shah, Ow-detr: open-world detection transformer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9235–9244.
[89] S. Ma, Y. Wang, Y. Wei, J. Fan, T.H. Li, H. Liu, F. Lv, Cat: localization and identification cascade detection transformer for open-world object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19681–19690.
[90] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: International Conference on Learning Representations, 2019, https://openreview.net/forum?id=Bkg6RiCqY7.
[91] J.-H. Kim, J. Jun, B.-T. Zhang, Bilinear attention networks, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems, vol. 31, Curran Associates, Inc., 2018, https://proceedings.neurips.cc/paper/2018/file/96ea64f3a1aa2fd00c72faacf0cb8ac9-Paper.pdf.
[92] J. Lu, J. Yang, D. Batra, D. Parikh, Hierarchical question-image co-attention for visual question answering, in: D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, R. Garnett (Eds.), Advances in Neural Information Processing Systems, vol. 29, Curran Associates, Inc., 2016, https://proceedings.neurips.cc/paper/2016/file/9dcb88e0137649590b755372b040afad-Paper.pdf.
[93] Y. Han, J. Yin, J. Wu, Y. Wei, L. Nie, Semantic-aware modular capsule routing for visual question answering, arXiv:2207.10404, 2022.
[94] M. Narasimhan, A.G. Schwing, Straight to the facts: learning knowledge base retrieval for factual visual question answering, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018.
[95] L. Liu, M. Wang, X. He, L. Qing, H. Chen, Fact-based visual question answering via dual-process system, Knowl.-Based Syst. 237 (2022) 107650, https://doi.org/10.1016/j.knosys.2021.107650, https://www.sciencedirect.com/science/article/pii/S0950705121009126.
[96] L. Zhang, S. Liu, D. Liu, P. Zeng, X. Li, J. Song, L. Gao, Rich visual knowledge-based augmentation network for visual question answering, IEEE Trans. Neural Netw. Learn. Syst. 32 (10) (2021) 4362–4373, https://doi.org/10.1109/TNNLS.2020.3017530.
[97] Y. Zhang, M. Jiang, Q. Zhao, Query and attention augmentation for knowledge-based explainable reasoning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15576–15585.
[98] M. Narasimhan, S. Lazebnik, A. Schwing, Out of the box: reasoning with graph convolution nets for factual visual question answering, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems, vol. 31, Curran Associates, Inc., 2018, https://proceedings.neurips.cc/paper/2018/file/c26820b8a4c1b3c2aa868d6d57e14a79-Paper.pdf.
[99] L. Song, J. Li, J. Liu, Y. Yang, X. Shang, M. Sun, Answering knowledge-based visual questions via the exploration of question purpose, Pattern Recognit. 133 (2023) 109015, https://doi.org/10.1016/j.patcog.2022.109015, https://www.sciencedirect.com/science/article/pii/S0031320322004952.
[100] J. Yu, Z. Zhu, Y. Wang, W. Zhang, Y. Hu, J. Tan, Cross-modal knowledge reasoning for knowledge-based visual question answering, Pattern Recognit. 108 (2020) 107563, https://doi.org/10.1016/j.patcog.2020.107563, https://www.sciencedirect.com/science/article/pii/S0031320320303666.
[101] M. Li, M. Marie-Francine, Dynamic Key-Value Memory Enhanced Multi-Step Graph Reasoning for Knowledge-Based Visual Question Answering, Association for the Advancement of Artificial Intelligence, 2021.
[102] A. Graves, S. Fernández, J. Schmidhuber, Bidirectional lstm networks for improved phoneme classification and recognition, in: W. Duch, J. Kacprzyk, E. Oja, S. Zadrożny (Eds.), Artificial Neural Networks: Formal Models and Their Applications – ICANN 2005, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005, pp. 799–804.
[103] S. Sukhbaatar, a. szlam, J. Weston, R. Fergus, End-to-end memory networks, in: C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, R. Garnett (Eds.), Advances in Neural Information Processing Systems, vol. 28, Curran Associates, Inc., 2015, https://proceedings.neurips.cc/paper/2015/file/8fb21ee7a2207526da55a679f0332de2-Paper.pdf, 2015.
[104] Y. Li, R. Zemel, M. Brockschmidt, D. Tarlow, Gated graph sequence neural networks, in: Proceedings of ICLR'16, Proceedings of iclr'16 Edition, 2016, https://www.microsoft.com/en-us/research/publication/gated-graph-sequence-neural-networks/.
[105] D. Garcia-Olano, Y. Onoe, J. Ghosh, Improving and diagnosing knowledge-based visual question answering via entity enhanced knowledge injection, in: Companion Proceedings of the Web Conference 2022, WWW '22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 705–715.
[106] E.-S. Kim, W.Y. Kang, K.-W. On, Y.-J. Heo, B.-T. Zhang, Hypergraph attention networks for multimodal learning, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14569–14578.

[107] Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, L.D. Jackel, Backpropagation applied to handwritten zip code recognition, Neural Comput. 1 (4) (1989) 541–551, https://doi.org/10.1162/neco.1989.1.4.541.
[108] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C.L. Zitnick, D. Parikh, Vqa: visual question answering, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015.
[109] K. Marino, X. Chen, D. Parikh, A. Gupta, M. 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 (CVPR), 2021, pp. 14111–14121.
[110] J. Wu, J. Lu, A. Sabharwal, R. Mottaghi, Multi-modal answer validation for knowledge-based vqa, Proc. AAAI Conf. Artif. Intell. 36 (3) (2022) 2712–2721, https://doi.org/10.1609/aaai.v36i3.20174, https://ojs.aaai.org/index.php/AAAI/article/view/20174.
[111] M. Luo, Y. Zeng, P. Banerjee, C. Baral, Weakly-supervised visual-retriever-reader for knowledge-based question answering, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 6417–6431, https://aclanthology.org/2021.emnlp-main.517.
[112] X. Chen, X. Wang, S. Changpinyo, A. Piergiovanni, P. Padlewski, D. Salz, S. Goodman, A. Grycner, B. Mustafa, L. Beyer, A. Kolesnikov, J. Puigcerver, N. Ding, K. Rong, H. Akbari, G. Mishra, L. Xue, A. Thapliyal, J. Bradbury, W. Kuo, M. Seyedhosseini, C. Jia, B. Karagol Ayan, C. Riquelme, A. Steiner, A. Angelova, X. Zhai, N. Houlsby, R. Soricut, PaLI: a jointly-scaled multilingual language-image model, arXiv:2209.06794, 2022.
[113] Z. Yang, Z. Gan, J. Wang, X. Hu, Y. Lu, Z. Liu, L. Wang, An empirical study of gpt-3 for few-shot knowledge-based vqa, Proc. AAAI Conf. Artif. Intell. 36 (3) (2022) 3081–3089, https://doi.org/10.1609/aaai.v36i3.20215, https://ojs.aaai.org/index.php/AAAI/article/view/20215.
[114] Y. Hao, H. Song, L. Dong, S. Huang, Z. Chi, W. Wang, S. Ma, F. Wei, Language models are general-purpose interfaces, arXiv:2206.06336, 2022.
[115] Y. Hu, H. Hua, Z. Yang, W. Shi, N.A. Smith, J. Luo, PromptCap: prompt-guided task-aware image captioning, arXiv:2211.09699, 2022.
[116] Y. Jiang, V. Natarajan, X. Chen, M. Rohrbach, D. Batra, D. Parikh, Pythia v0.1: the winning entry to the VQA challenge 2018, arXiv:1807.09956, 2018.
[117] A. Kamath, C. Clark, T. Gupta, E. Kolve, D. Hoiem, A. Kembhavi, Webly supervised concept expansion for general purpose vision models, arXiv:2202.02317, 2022.
[118] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, H. Yang, OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework, in: K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, S. Sabato (Eds.), Proceedings of the 39th International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 162, PMLR, 2022, pp. 23318–23340, https://proceedings.mlr.press/v162/wang22al.html.
[119] E. Perez, F. Strub, H. de Vries, V. Dumoulin, A. Courville, Film: visual reasoning with a general conditioning layer, Proc. AAAI Conf. Artif. Intell. 32 (1) (Apr. 2018), https://doi.org/10.1609/aaai.v32i1.11671, https://ojs.aaai.org/index.php/AAAI/article/view/11671.
[120] Z. Yu, J. Yu, C. Xiang, J. Fan, D. Tao, Beyond bilinear: generalized multimodal factorized high-order pooling for visual question answering, IEEE Trans. Neural Netw. Learn. Syst. 29 (12) (2018) 5947–5959, https://doi.org/10.1109/TNNLS.2018.2817340.
[121] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, Bottom-up and top-down attention for image captioning and visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[122] Z. Yu, J. Yu, Y. Cui, D. Tao, Q. Tian, Deep modular co-attention networks for visual question answering, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
[123] Q. Cao, B. Li, X. Liang, L. Lin, Explainable high-order visual question reasoning: a new benchmark and knowledge-routed network, arXiv:1909.10128, 2019.
[124] S. Shen, L.H. Li, H. Tan, M. Bansal, A. Rohrbach, K.-W. Chang, Z. Yao, K. Keutzer, How much can CLIP benefit vision-and-language tasks?, in: International Conference on Learning Representations, 2022, https://openreview.net/forum?id=zf_Ll3HZWgy.
[125] J. Li, D. Li, S. Savarese, S. Hoi, Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models, in: Proceedings of the 40th International Conference on Machine Learning, ICML'23, JMLR.org, 2023.
[126] O.J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F.L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Baltescu, H. Bao, M. Bavarian, J. Belgum, I. Bello, J. Berdine, G. Bernadett-Shapiro, C. Berner, L. Bogdonoff, O. Boiko, M. Boyd, A.-L. Brakman, G. Brockman, T. Brooks, M. Brundage, K. Button, T. Cai, R. Campbell, A. Cann, B. Carey, C. Carlson, R. Carmichael, B. Chan, C. Chang, F. Chantzis, D. Chen, S. Chen, R. Chen, J. Chen, M. Chen, B. Chess, C. Cho, C. Chu, H.W. Chung, D. Cummings, J. Currier, Y. Dai, C. Decareaux, T. Degry, N. Deutsch, D. Deville, A. Dhar, D. Dohan, S. Dowling, S. Dunning, A. Ecoffet, A. Eleti, T. Eloundou, D. Farhi, L. Fedus, N. Felix, S.P. Fishman, J. Forte, I. Fulford, L. Gao, E. Georges, C. Gibson, V. Goel, T. Gogineni, G. Goh, R. Gontijo-Lopes, J. Gordon, M. Grafstein, S. Gray, R. Greene, J. Gross, S.S. Gu, Y. Guo, C. Hallacy, J. Han, J. Harris, Y. He, M. Heaton, J. Heidecke, C. Hesse, A. Hickey, W. Hickey, P. Hoeschele, B. Houghton, K. Hsu, S. Hu, X. Hu, J. Huizinga, S. Jain, S. Jain, J. Jang, A. Jiang, R. Jiang, H. Jin, D. Jin, S. Jomoto, B. Jonn, H. Jun, T. Kaftan, L. Kaiser, A. Kamali, I. Kanitscheider, N.S. Keskar, T. Khan, L. Kilpatrick, J.W. Kim, C. Kim, Y. Kim, H. Kirchner, J.R. Kiros, M. Knight, D. Kokotajlo, L. Kondraciuk, A. Kondrich, A. Konstantinidis, K. Kosic, G. Krueger, V. Kuo, M. Lampe, I. Lan, T. Lee, J. Leike, J. Leung, D. Levy, C.M. Li, R. Lim, M. Lin, S. Lin, M. Litwin, T. Lopez, R. Lowe, P. Lue, A.A. Makanju, K. Malfacini, S. Manning, T. Markov, Y. Markovski, B. Martin, K. Mayer, A. Mayne, B. McGrew, S.M. McKinney, C. McLeavey, P. McMillan, J. McNeil, D. Medina, A. Mehta, J. Menick, L. Metz, A. Mishchenko, P. Mishkin, V. Monaco, E. Morikawa, D.P. Mossing, T. Mu, M. Murati, O. Murk, D. M'ely, A. Nair, R. Nakano, R. Nayak, A. Neelakantan, R. Ngo, H. Noh, O. Long, C. O'Keefe, J.W. Pachocki, A. Paino, J. Palermo, A. Pantuliano, G. Parascandolo, J. Parish, E. Parparita, A. Passos, M. Pavlov, A. Peng, A. Perelman, F. de Avila Belbute Peres, M. Petrov, H.P. de Oliveira Pinto, M. Pokorny, M. Pokrass, V.H. Pong, T. Powell, A. Power, B. Power, E. Proehl, R. Puri, A. Radford, J. Rae, A. Ramesh, C. Raymond, F. Real, K. Rimbach, C. Ross, B. Rotsted, H. Roussez, N. Ryder, M.D. Saltarelli, T. Sanders, S. Santurkar, G. Sastry, H. Schmidt, D. Schnurr, J. Schulman, D. Selsam, K. Sheppard, T. Sherbakov, J. Shieh, S. Shoker, P. Shyam, S. Sidor, E. Sigler, M. Simens, J. Sitkin, K. Slama, I. Sohl, B.D. Sokolowsky, Y. Song, N. Staudacher, F.P. Such, N. Summers, I. Sutskever, J. Tang, N.A. Tezak, M. Thompson, P. Tillet, A. Tootoonchian, E. Tseng, P. Tuggle, N. Turley, J. Tworek, J.F.C. Uribe, A. Vallone, A. Vijayvergiya, C. Voss, C. Wainwright, J.J. Wang, A. Wang, B. Wang, J. Ward, J. Wei, C. Weinmann, A. Welihinda, P. Welinder, J. Weng, L. Weng, M. Wiethoff, D. Willner, C. Winter, S. Wolrich, H. Wong, L. Workman, S. Wu, J. Wu, M. Wu, K. Xiao, T. Xu, S. Yoo, K. Yu, Q. Yuan, W. Zaremba, R. Zellers, C. Zhang, M. Zhang, S. Zhao, T. Zheng, J. Zhuang, W. Zhuk, B. Zoph, Gpt-4 technical report, 2023, https://api.semanticscholar.org/CorpusID:257532815.
[127] N. Balepur, A. Ravichander, R. Rudinger, Artifacts or abduction: How do llms answer multiple-choice questions without the question?, preprint, 2024.
[128] B. Smith, Stop talking about tomorrow's ai doomsday when ai poses risks today, Nature 618 (2023) 885–886.
[129] P. Samuelson, Generative ai meets copyright, Science 381 (6654) (2023) 158–161, https://doi.org/10.1126/science.adi0656, https://www.science.org/doi/pdf/10.1126/science.adi0656, https://www.science.org/doi/abs/10.1126/science.adi0656.
[130] A. Radhakrishnan, D. Beaglehole, P. Pandit, M. Belkin, Mechanism for feature learning in neural networks and backpropagation-free machine learning models, Science 383 (6690) (2024) 1461–1467, https://doi.org/10.1126/science.adi5639, https://www.science.org/doi/pdf/10.1126/science.adi5639, https://www.science.org/doi/abs/10.1126/science.adi5639.
[131] R. Mottaghi, M. Rastegari, A. Gupta, A. Farhadi, "what happens if..." learning to predict the effect of forces in images, in: B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.), Computer Vision – ECCV 2016, Springer International Publishing, Cham, 2016, pp. 269–285.
[132] H. Gu, et al., The Discourses and Sayings of Confucius: A New Special Translation, Illustrated with Quotations from Goethe and Other Writers, Kelly and Walsh, limited, 1898.
[133] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, A. Kalyan, Learn to explain: multimodal reasoning via thought chains for science question answering, in: A.H. Oh, A. Agarwal, D. Belgrave, K. Cho (Eds.), Advances in Neural Information Processing Systems, 2022, https://openreview.net/forum?id=HjwK-Tc_Bc.
[134] P. Gao, Z. Jiang, H. You, P. Lu, S.C.H. Hoi, X. Wang, H. Li, Dynamic fusion with intra- and inter-modality attention flow for visual question answering, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[135] P. Lu, L. Qiu, J. Chen, T. Xia, Y. Zhao, W. Zhang, Z. Yu, X. Liang, S.-C. Zhu, Iconqa: a new benchmark for abstract diagram understanding and visual language reasoning, in: The 35th Conference on Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks, 2021.
[136] L.H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, K.-W. Chang, What does BERT with vision look at?, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 5265–5275, Online, https://aclanthology.org/2020.acl-main.469.
[137] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P.J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, J. Mach. Learn. Res. 21 (140) (2020) 1–67, http://jmlr.org/papers/v21/20-074.html.
[138] A. Ben Abacha, M. Sarrouti, D. Demner-Fushman, S.A. Hasan, H. Müller, Overview of the vqa-med task at imageclef 2021: visual question answering and generation in the medical domain, in: CLEF 2021 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bucharest, Romania, 2021.
[139] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02, Association for Computational Linguistics, USA, 2002, pp. 311–318.
[140] H. Gong, R. Huang, G. Chen, G. Li, Sysu-hcp at vqa-med 2021: a data-centric model with efficient training methodology for medical visual question answering, in: CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania, CEUR Workshop Proceedings, 2021.
[141] Q. Xiao, X. Zhou, Yunnan university at vqa-med 2021: pretrained biobert for medical domain visual question answering, in: CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania, CEUR Workshop Proceedings, 2021.
[142] S. Eslami, G. de Melo, C. Meinel, Teams at vqa-med 2021: Bbn-orchestra for long-tailed medical visual question answering, in: CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania, CEUR Workshop Proceedings, 2021.
[143] J. Li, S. Liu, Lijie at imageclefmed vqa-med 2021: attention model-based efficient interaction between multimodality, in: CLEF (Working Notes), 2021, pp. 1275–1284.
[144] R. Schilling, P. Messina, D. Parra, H. Löbel, Puc chile team at vqa-med 2021: approaching vqa as a classification task via fine-tuning a pretrained cnn, in: CLEF (Working Notes), 2021, pp. 1346–1351.
[145] Y. Li, Z. Yang, T. Hao, Tam at vqa-med 2021: a hybrid model with feature extraction and fusion for medical visual question answering, in: CLEF (Working Notes), 2021, pp. 1295–1304.
[146] N.M.S. Sitara, K. Srinivasan, Ssn mlrg at vqa-med 2021: an approach for vqa to solve abnormality related queries using improved datasets, in: CLEF (Working Notes), 2021, pp. 1329–1335.