ITC 3/54
Information Technology and Control
Vol. 54 / No. 3/ 2025
pp. 751-767
DOI 10.5755/j01.itc.54.3.41097

Task-guided Visual Information Extracting Network for Visual Question Answering

Received 2025/04/07 | Accepted after revision 2025/06/10

HOW TO CITE: Cong, Y., Mo, H. (2025). Task-guided Visual Information Extracting Network for Visual Question Answering. *Information Technology and Control*, 54(3), 751-767. https://doi.org/10.5755/j01.itc.54.3.41097

Task-guided Visual Information Extracting Network for Visual Question Answering

Yao Cong, Hongwei Mo

College of Intelligent Systems Science and Engineering, Harbin Engineering University, 145 Nantong Street, Nangang District, Harbin, P. R. China; e-mail: honwei2004 @126.com

Corresponding author: honwei2004 @126.com

The most expected approach to Visual Question Answering tasks is to observe the scene and answer questions with human-like reasoning. Early multimodal fusion schemes focused more on the final result rather than the intermediate reasoning process. In contrast, the step-by-step reasoning method based on task decomposition is more capable of meeting the visual reasoning requirements in the task. Nevertheless, the real-world performance of most current models based on step-by-step reasoning is inadequate due to the absence of required reasoning information and the incapability to generate appropriate solution approaches when confronting real scenes and natural language questions presented by humans. A VQA model based on scene information extraction network (SIEN-VQA) is proposed to address the above issues. SIEN-VQA utilizes graph structured data and Task Decomposition Network to generate reasoning steps, extract relevant image scene information based on the reasoning steps for reasoning execution, and enhances the model's reasoning execution ability in natural language and real scenes. We conducted experimental validation on the CLEVR-Human and GQA datasets, and the validation results showed that our model is able to decompose and conquer problems according to human-like logic, and extract effective scene information that is relevant to the task, which improves the accuracy of answering questions compared to the comparative model.

KEYWORDS: Visual question answering, scene information, visual reasoning, step-by-step reasoning, SIEN-VQA

1. Introduction

Visual Question Answering (VQA) is a comprehensive machine vision task that has received widespread attention, the origin of question answering (QA) task can be traced back to the Turing test, the VQA task was first proposed in 2014 and has since then attracted a lot of academic attention, especially after the relevant dataset was proposed (DAQUAR, COCO-QA, VQA v2.0, etc.), in contrast with QA tasks, the com-

plex scene analysis has been added in VQA tasks due to the necessity of obtaining and processing pertinent information from visual scenes as well as understanding natural language [1], [19]. Thus, VQA tasks involve natural language understanding, scene understanding, and visual reasoning, this task can be applied to various intelligent assistance fields, lies at the core of developing embodied intelligence in robotic systems as well [5], [26]. Especially when combined with robots, by bridging multimodal comprehension and real-world interaction capabilities, VQA enables robots to dynamically adapt to environmental contexts and perform complex tasks—a fundamental requirement for achieving true embodied intelligence. The early solution for VQA tasks is generally multimodal feature fusion, which involves encoding and fusing natural language and image features, feeding the fused features into neural network for training, and selecting answers based on probability [34]. Considering the large amount of irrelevant information in the fusion process, scholars have adopted attention mechanisms to limit it, thereby improving the accuracy of answer selection. As a whole, the multimodal fusion scheme focuses more on the results of answering questions and neglects the process of visual reasoning [10], [51]. Visual reasoning aims to enable machines not only to recognize objects in images but also to understand logical relationships within scenes, infer implicit information, and make complex decisions [2]. Intermediate reasoning processes are critical to the logical integrity of visual reasoning. On the one hand, intermediate steps make the decision-making path visible, enhancing the interpretability of the model for human understanding [39]. On the other hand, models that emphasize reasoning processes are better equipped to handle more sophisticated reasoning tasks. In practical applications, explicitly defining intermediate reasoning steps facilitates fault diagnosis and error correction, ensuring robustness and reliability in real-world scenarios. Andreas et al. [3], [4] proposed a modular network solution based on task decomposition, which essentially decomposes the overall natural language problem into a combination of multiple subtasks. By completing the subtasks, the overall task is completed to obtain the final answer. This approach places more emphasis on the reasoning process and has attracted the attention of most scholars since it was proposed. Especially when Stanford researchers proposed the CLEVR dataset, a large number of scholars began to focus on the visual reasoning process involved in VQA. In 2022, Google introduced the Chain-of-Thought (CoT) methodology for question-answering tasks, the core innovation of CoT lies in prompting large language models (LLMs) to generate not just final answers but also the explicit reasoning pathways leading to those conclusions [45]. This paradigm deliberately mimics human cognitive processes by explicitly simulating the sequential reasoning patterns observed in human problem-solving. The demonstrated effectiveness of CoT in improving answer accuracy reinforces the premise that making intermediate reasoning steps explicit enhances computational reasoning capabilities, particularly when handling complex multi-step problems that require systematic information synthesis, which aligning closely with the fundamental principles of task-decomposition-based Visual Question Answering (VQA) approaches, both bridge perception and cognition through structured intermediate representations (such as natural language steps or symbolic logic), avoiding the "black box decision-making" of end-to-end models.

At present, three elements can be summarized from the VQA scheme based on task decomposition, including relevant clue acquisition, reasoning step generation, and reasoning step execution. Each element will affect the final reasoning result. Currently, visual question answering models based on task decomposition mainly use scene graphs, sequence-to-sequence neural networks, and a set of meta operation functions to complete the three elements, and verify them on a synthetic dataset. Therefore, when facing with real scenes and natural language problems raised by humans, problems often arise where relevant clues are not fully obtained and not adaptable to the flexibility of natural language to generate correct reasoning steps, which affects the execution of reasoning. In this regard, this work proposes a step-by-step neural-symbolic reasoning method based on scene information extraction for VQA tasks, which improves the reasoning execution ability by improving the accuracy of reasoning step generation and the completeness of task related scene information. We observed that the visual cues required to address a specific problem are often implicitly indicated within the question prompt. To leverage this, we analyze the question and utilize its inherent constraints to delineate the scope of relevant visual information, thereby minimizing interference from redundant data—a methodology we

define as task-guided visual information extraction.

Specifically, we have done the following work: We constructed a Task Decomposition Network for reasoning steps generation, using a graph-to-sequence structured network, and trained with graph structured data to improve the network's adaptability to natural languages, in addition, we constructed a Scene Information Extraction Network for scene information extraction, and using the generated reasoning steps as a kind of CoT-like prompt to obtain the most relevant reasoning information to the task and thus improve the reasoning execution ability.

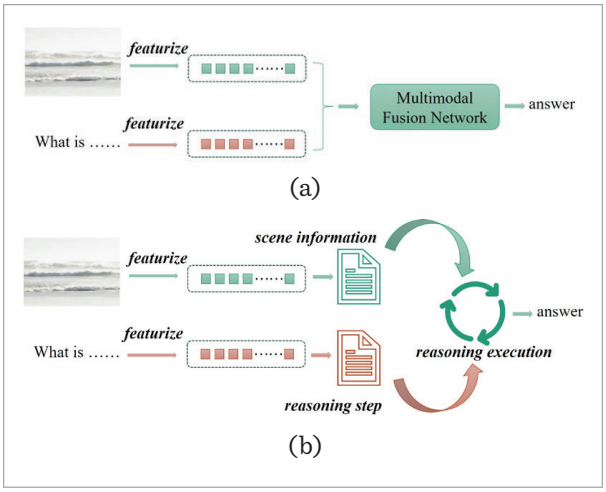
2. Related Work

Prior to the deep learning revolution, visual reasoning predominantly relied on handcrafted feature descriptors (SIFT, HOG, SURF) to encode low-level visual patterns—color histograms, edge orientations, and texture gradients. These methods enabled localized analysis of scene elements but struggled to capture semantic relationships or contextual dynamics. The advent of CNN-based architectures marked a paradigm shift, achieving perceptual-level visual inference through foundational tasks like image classification (object categorization), bounding box detection (spatial localization), pixel-wise segmentation (fine-grained scene parsing). While effective in extracting atomic features (object types, geometric attributes, motion trajectories), these models operated as isolated perception engines—they lacked mechanisms for cognitive synthesis. True scene understanding necessitates transcending perception to model including inter-object relations (spatial, functional, or causal dependencies), event dynamics (temporal causality between actions), contextual grounding (integrating domain-specific commonsense). The emergence of Visual Question Answering (VQA) introduced multimodal paradigms, merging language queries with visual inputs to drive systematic reasoning. This transition from unimodal to cross-modal processing addressed critical limitations. The VQA task was initially proposed by Malinowski et al. [32] as a new task to gauge the level of understanding of the scene. Addressing VQA tasks necessitates the featurization and the joint comprehension for both image and text, consequently, there have been extensive research on these aspects, leading to the development of VQA systems having diverse algorithms. In this section, we will introduce these VQA methods into two categories in relation to the answer generation procedure.

2.1. One-stage VQA Method based on End-to-End Model

After proposing the VQA task, Malinowski et al. [33] approached it as a problem of solving for the posterior probability $\mathbf{P}(A|Q, S)$ of answer A when presented with question Q and scene segmentation S. They subsequently proposed the Neural-Image-QA model [33], which analyzes images using convolutional neural networks (CNN) and inputs the question together with visual representations into the Long Short-Term Memory (LSTM) network. Similarly, Ren et al. [37] proposed the VIS+LSTM model, they employed a linear or affine transformation to align the dimension between image feature vectors and word embeddings, then treated images as one word of the question, together with question text handed over to the LSTM network for processing. Fukui et al. [12] applied Multimodal Compact Bilinear pools the extracted multimodal feature vectors, and obtained answers by considering it as multi-class classification problem with 3000 possible categories. Ben-younes et al. [6] used a fully convolutional neural network to describe the image content, and a GRU recurrent network for the question, then fusing them via Tucker decomposition. These VQA system directly generates answers through an end-to-end network. The entire process can be summarized as follows: extracting feature vector from the image and the question, and send them into the end-to-end network to fuse the two modalities to generate the answer. The key to this one-stage generation method lies in how to fuse multimodal features to achieve joint comprehension, hence a large number of studies are carried out with this topic as the center. The commonality of all studies lies in extracting corresponding features from text and images through neural networks such as CNN and LSTM before fusion, and mapping the features to a specified dimension [53], [23]. Typical multimodal feature fusion methods mainly combine visual features and textual features through concatenation [18], [49], [50], element-wise addition [14], [30] and element-wise multiplication [44], [52]. The research found that in the process of feature fusion, both visual features and text features have elements that are not related to answering questions.

Figure 1
One-stage VQA method (at the top) and Multi-stage VQA method (at the bottom).



Considering this, in order to further improve the accuracy of answering questions, researchers have optimized the VQA model by using attention mechanism to increase the importance of useful information in visual and text information. For example, Zhu et al. [55] combined attention methods with LSTM network. At the encoding stage, taking the image as the first input token, taking image feature together with text feature as input then outputting the attention map for each step, and multiplying the attention map with visual features to generate new visual features. Shih et al. [40] directly multiplied visual features with text features to obtain attention weights, and the size of weights represents the importance of the regions. The improvement effect of attention mechanism on multimodal fusion models lies in aligning image information with problem information. Therefore, more changes have been made to the attention mechanism aimed at improving the alignment effect. Lu et al. [31] utilized co-attention to jointly learn the attention of images and texts, proposed parallel co-attention and alternating co-attention that differ in the order in which image and question attention maps are generated. Gao et al. [13] proposed a dynamic fusion model with intra modal and inter modal attention. This method dynamically uses the inter modal learns the cross-modal interactions between the image regions and question words and uses intra modal to model word-to-word relations and region-to-region relations. Yu et al. [48] proposed a visual question answering system MLAN based on a multi-level attention network. The attention in this system includes semantic attention, attribute attention, and visual attention. The model focuses on semantic attributes and image regions related to the problem, narrowing the semantic gap between vision and language. After the addition of attention mechanism, the accuracy of VQA models has indeed improved, but the overall process of answering questions in the model is still a black box model that is unclear and lacks interpretability. Scholars have begun to consider whether these VQA models rely on data bias in the process of answering questions, rather than actually reasoning. How to establish an effective and interpretable universal model has always been a hot academic research topic. Nowadays, vision-language pre-training models based on self-attention architectures, such as ALBEF [28], InstructBLIP [9], etc. are becoming more popular and these multimodal fusion large models have achieved remarkable results in the VQA field. Moreover, with the rise of ChatGPT, LLM has attracted widespread attention. These models were trained on a large amount of data through prompt learning and reinforcement learning from human feedback, some of vision-language models was developed based on LLM's capacity of facilitating natural language communication that leverages their distinct and complementary capabilities, such as MiniGPT-4 [54], Cola [7] etc. However, both LLM and multimodal fusion large models still have limitations in its visual reasoning ability especially when facing multi-step reasoning problems.

2.2. Multi-stage VQA Method based on Task Decomposition

We would prefer the machine to answer these complex questions based on human logic. Andreas was inspired by the fact that humans divide questions into several steps when answering them, and serialized the reasoning process of visual question answering models. The NMN [3],[4] (Neural Module Networks) model constructed multiple sub-modules required for answering questions, such as "find", "transform", "combine", "describe", "measure", etc. Afterwards, according to the structure of the problem, the combination of sub-modules is automatically generated for collaborative learning. This method is significantly different from the one-stage method based on multimodal mode fusion, as shown

in the Figure.1, it is a multi-stage method that begins by analyzing the reasoning steps of the question, and then combines image information to reach the final result step by step.

In 2016, Stanford and Facebook jointly released the CLEVR dataset [25], which is called the Composite Language and Elementary Visual Reasoning diagnostic dataset, the dataset carefully controls the potential bias and tests a range of visual reasoning abilities, research has shown that many VQA methods based on modal fusion exhibit a significant decrease in performance on CLEVR dataset. In 2017, Johnson et al. [24] proposed a new modular visual reasoning method for the CLEVR dataset, they argue that to successfully perform complex reasoning tasks, it might be necessary to explicitly incorporate compositional reasoning in the model structure. Due to the simple scene of the dataset, only need to pay attention on the reasoning itself. As a result, researches on this composite dataset and VQA method based on task decomposition have grown rapidly and mainly focuses on NMN family, such as Hu et al. [15], [17] proposed an end-to-end modular network (N2NMN) that can directly predict the module combination and layout of a problem without the help of a parser. However, both NMN and N2NMN require strong supervised information to pretrain or supervise layout strategies to obtain correct module layout and maintain good performance. If these supervised signals are lost, the model will experience significant performance degradation or inability to converge. Based on this, Hu et al. [16] proposed a Stack-NMN method that automatically guides the required sub-task decomposition for combinatorial reasoning without relying on strong supervised signals. At the same time, it can also ensure that the layout strategy of the module is differentiable, allowing for optimization using gradient descent method. Yamada et al. [46] proposed Transformer Module Network (TMN), it is a kind of NMN based on compositions of Transformer modules, combining the strengths of Transformers and NMNs in order to improve the systematic generalization capabilities of learning machines. Yi et al. [47] proposed Neural-Symbolic VQA (NS-VQA), which is a typical model composed of scene parser, problem parser, and program executors, laying the foundation for later visual reasoning models, compared to NMN, NS-VQA mainly simplifies the execution of sub-tasks. Eiter et al. [11] presented a neuro-symbolic VQA pipeline for CLEVR, it relies on answer-set programming (ASP) to infer the right answer given the neural network output and a confidence threshold. The necessary stages mainly include object detection, confidence thresholds, ASP encoding, the distinguishing feature of the pipeline is fixing the threshold based on the mean and the standard deviation of prediction scores so that restricting non-determinism of object detection prediction, meanwhile, ASP offers a simple yet expressive modelling language and efficient solver technology. In Table 1, we summarize the advantages and disadvantages of one-stage and multi-stage VQA methods. As shown in the comparison, multi-stage VQA methods place greater emphasis on reasoning processes and demonstrate better adaptability to hierarchical problems. In contrast, one-stage VQA methods are more suitable for tasks requiring real-time responses or those with low dependency on intermediate reasoning steps. For complex reasoning tasks, multi-stage VQA methods exhibit significant advantages due to their structured approach.

Table 1
Comparison of one-stage and multi-stage VQA methods.

VQA method	Advantages	Disadvantages
One-stage	Computational efficiency, end-to-end optimization, reduced error propagation.	Potential over-simplification of complex reasoning chains, limited interpretability.
multi-stage	Explicit modeling of intermediate logic, better error diagnosis, adaptability to hierarchical problems.	Higher computational cost, risk of error accumulation across stages.

For VQA, the key objective is to acquire the reasoning steps, namely sub-task decomposition, which is a characteristic of task decomposition VQA system. The generation of reasoning steps is the prerequisite for everything, followed by precise execution of reasoning steps to obtain the expected result. The current VQA system built on task decomposition predominantly directs its attention to the training and testing of synthetic datasets, wherein the scene content is simplified and language templates are employed to produce structured text as questions. When faced with real scenes and human natural language, there is difficulty in generating correct reasoning steps as well as a lack of reasoning information to infer the correct answer to the problem.

In contrast to the other approaches, our work has the benefit of considering not only word semantic information but also implicit structural information in the language, graph structured data is much better at representing and utilizing the implicit structural information than serialized data, thus the Graph2Seq Task Decomposition Network learned from graph data can better adapt to non-normalized natural languages compared to Seq2Seq networks, improve the accuracy of VQA tasks by improving the generation effect of reasoning steps. Furthermore, we employ the generated reasoning steps as prompts to guide the construction of scene graphs, thereby reducing redundant information retrieval and enhancing task-specific information relevance. This approach ultimately improves the overall accuracy of VQA tasks—a framework we term task-guided visual information extracting.

3. Method

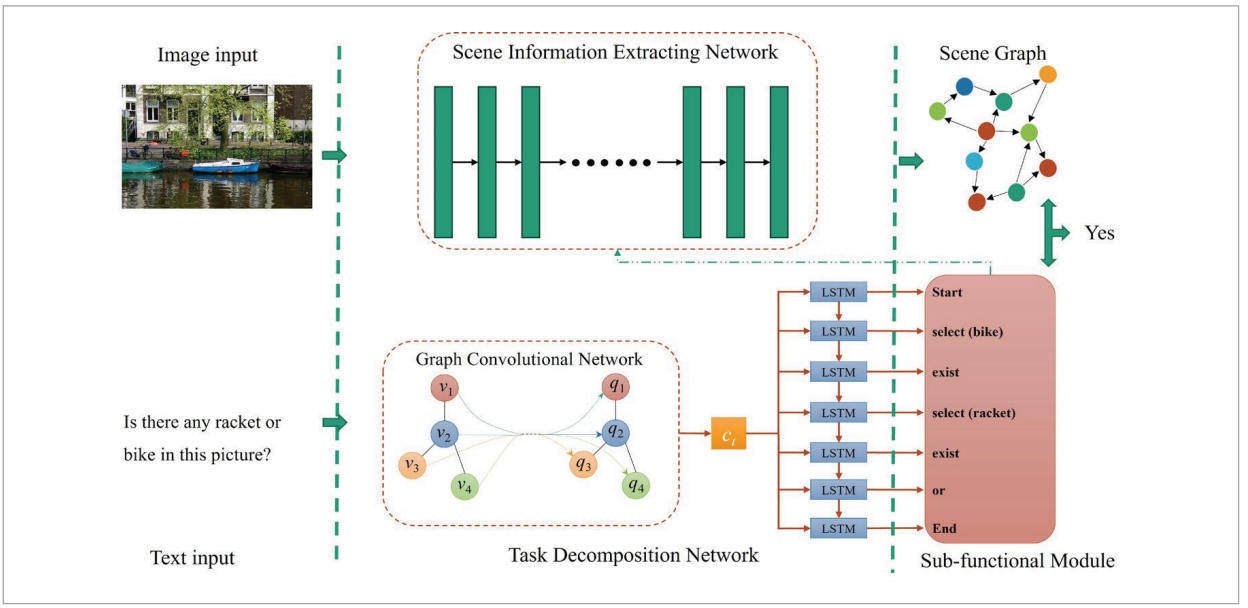
Our SIEN-VQA model has three components: the Task Decomposition Network (TDN), the Scene Information Extracting Network (SIEN) and a set of reasoning execution components, the entire model structure has shown in Figure 2. For the visual reasoning required in VQA tasks, it is necessary to clarify the goals and steps of the reasoning, obtain relevant visual clues, and through a series of tinny reasoning operations to get the final result. The task decomposition network parsing the entire reasoning task into multiple reasoning steps based on input text, the scene information extracting network is used to extract the visual information required for reasoning, and the reasoning execution component are a set of composable program modules that address specific sub-tasks and have reasoning capabilities. The relevant framework and principles will be further introduced in detail below.

3.1. Task Decomposition Network

In VQA tasks, the reasoning task is decomposed by parsing the reasoning steps from the natural language question and the graph structured data are employed to express natural language. Our data is more comprehensive than serialized data, as it contains syntactic structure information and other information. Therefore, we construct dependency graphs for question text inputs, and use GCN and LSTM to form an Encoder-Decoder network with Graph-to-Sequence structure to process graph structured data.

Figure 2

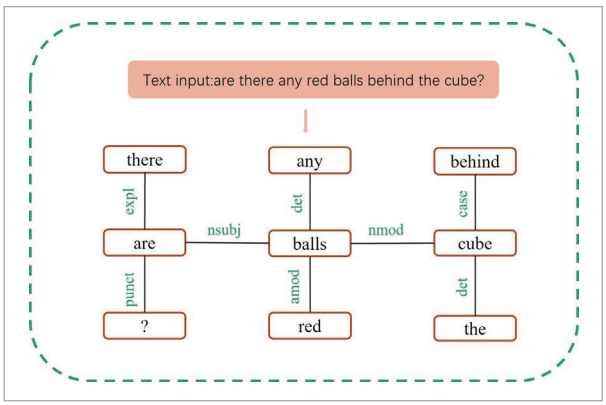
SIEN-VQA model. The model contains three components: the Task Decomposition Network (TDN), the Scene Information Extracting Network (SIEN) and a set of reasoning execution components which correspond to the three elements of generating inference steps, obtaining relevant clues, and executing inference steps, respectively.



Building an introduced graph refers to building the graph structure during preprocessing by leveraging existing relationship resolution tools or manually defined rules. Dependency graphs can provide prior structures to participate in parsing as introduced graphs. A dependency graph mainly describes the dependencies between different words in a given sentence, the dependency graph for the question "Are there any red balls behind the cube?", as parsed by Stanford NLP, is shown in Figure 3.

Figure 3

Dependency graph constructed by Stanford NLP.



As we can see, the dependency graph contains multiple edges representing dependencies between words. The entire dependency graph will be used as a heterogeneous introduced graph to train the task decomposition network. During this process, the edges carrying syntactic information will play an important role in the transmission and updating of messages. The process is as follows:

Node and edge embedding: For the input graph data G = (V, E), where V is the set of nodes and E is the set of edges, computing the adjacency matrix A and degree matrix D, and embedding the nodes and edges:

$$v_i^* = W_N v_i, e_{ij}^* = W_E e_{ij}, \tag{1}$$

where v_i represents the i-th node, v^* irepresents the embedded feature vector of the i-th node, e_{ij} represents the edge between node v_i and node v_j, e^*_{ij} represents the embedded feature vector of edge e_{ij}, WN and WE respectively represent the weight matrix for node and edge embedding. Then, the set of embedded node features is $\boldsymbol{V}^*$ and the set of embedded edge features is $\boldsymbol{E}^*$.

Encoding with GCN: Constructing a L-layer GCN with input graph feature data set $\boldsymbol{G}_f = \{\boldsymbol{A}, \boldsymbol{D}, \textbf{cat}\ (\boldsymbol{V}^*, \boldsymbol{E}^*)\}$, where **cat** represents the concatenation operation, and propagating and updating node and edge information layer by layer:

$$Q^{(l+1)} = \sigma\left(\Gamma Q^{(l)} W^{(l)} + b^{(l)}\right) \tag{2}$$

$$\Gamma = D^{-\frac{1}{2}}\left(D - A\right)D^{-\frac{1}{2}}, \tag{3}$$

where Q(l) represents the feature information of the l-th GCN layer, $W(l)$and $b(l)$ represent the weight matrix and bias matrix of the l-th GCN layer, σ denotes the nonlinear activation function, and Γ is the Laplacian matrix.

Decoding with LSTM: Obtaining the hidden state h_t at current time t through the LSTM network on the decoder side.

$$h_t = LSTM\left(h_{t-1}, y_{t-1}\right), \tag{4}$$

where h_{t-1} denotes the hidden state vector from the preceding timestep t-1, while y_{t-1} represents the predicted output symbol generated at the prior temporal iteration. The attention weight α is calculated from h_t:

$$\alpha_i = \frac{\exp\left(h_t^{\mathrm{T}} W_A q_i^{(L)}\right)}{\sum_{j=0}^{n} \exp\left(h_t^{\mathrm{T}} W_A q_j^{(L)}\right)}, \tag{5}$$

where, W_{A} is the attention weight matrix, $q(L)$ i is the i-th feature vector in Q(L), and α_i is the attention value for the i-th feature vector.

Calculating the context vector c_t at time t based on the data feature Q(L) calculated from the last GCN layer of the encoder and α:

$$c_t = \sum_i^n \alpha_i q_i^{(L)}. \tag{6}$$

Finally, the context vector, together with the decoder output, is passed to a fully connected layer with softmax activation to obtain the distribution for the predicted token:

$$p\left(y_t \mid h_t, c_t\right) = soft\max\left(h_t, c_t\right). \tag{7}$$

The overall loss function is as follows:

$$l_{cross} = -\sum_{t}^{N} y_t' \log p_t, \tag{8}$$

where y_t' is the true label of the t-th token, and p_t is the distribution for the t-th predicted token, N represents the length of the final symbol sequence output by the network.

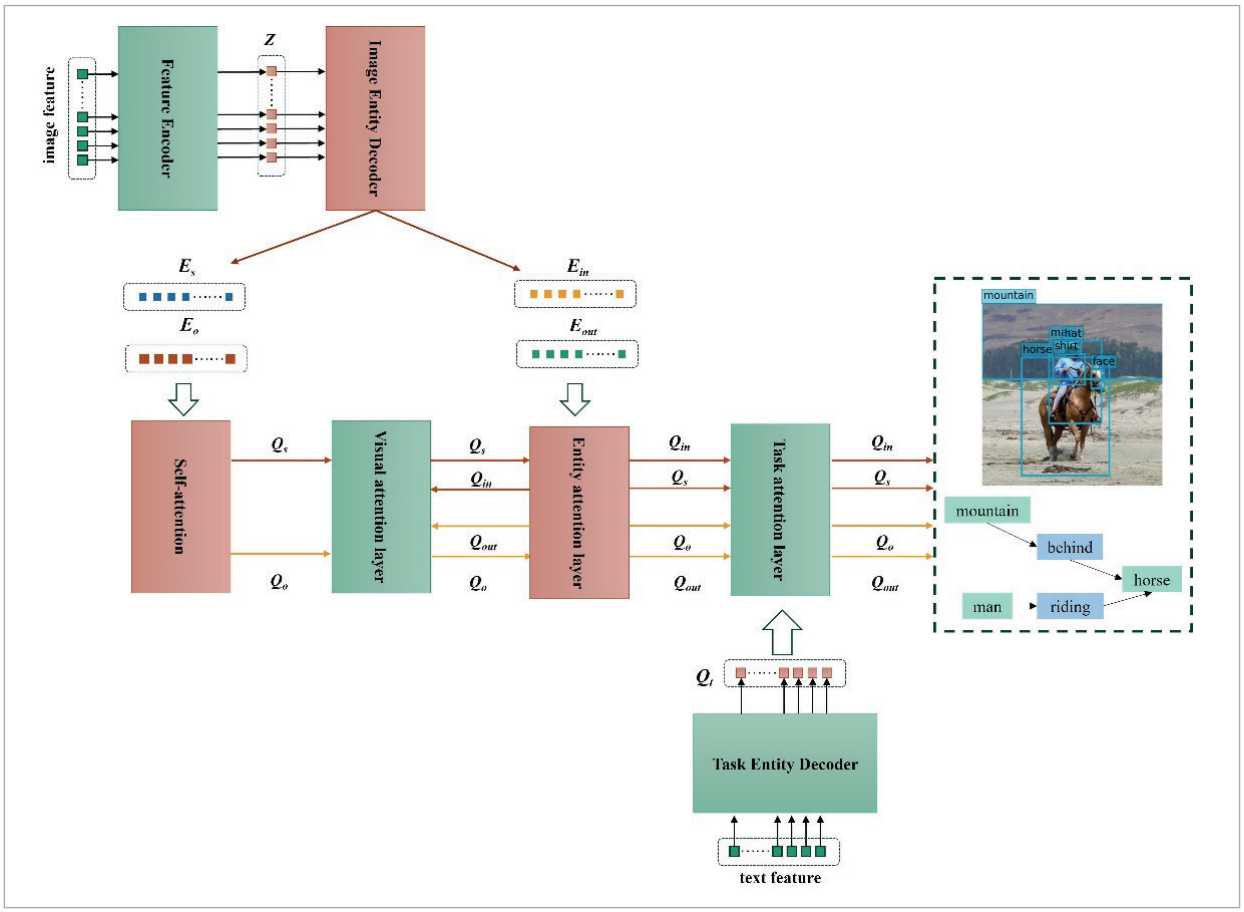
3.2. Scene Information Extracting Network

SIEN-VQA treats the VQA task as a reasoning task and analyze the scene information based on the generated reasoning steps to obtain the reasoning results. The scene information includes entity categories, entity attributes, and relationships between entities. The completeness of the scene information directly affects whether the reasoning steps can be executed accurately. The ideal situation is to be able to extract the most pertinent details to the reasoning task from the image, disregarding any inconsequential information in the image. Earlier approaches regularly employed object detection-based techniques to analyze images, i. e., object detection was performed initially followed by the further detection of the relationships between objects. This purposeless parsing may result in too much irrelevant information or a lack of key information. To address this issue, we construct a Scene Information Extraction Network, as shown in Figure 4.

We add the generation step as a prompt to the scene parsing process, thereby limiting the detection process. The specific implementation method is as follows:

To generate a scene graph, we first obtain the image feature encoding Z and a set of entity bounding boxes (n) B_I from the image I. Then, Z and B_I are fed into an image entaity decoder to obtain the subject encodings E_s, object encodings E_o, inbound edge encodings E_{in}, and outbound edge encodings E_{out}.

Figure 4
Scene Information Extracting Network.



By self-attention captures the context between triplets and the dependencies between all subjects and objects:

$$A_{score} = softmax\left(\frac{[E_s, E_o]W_Q\left([E_s, E_o]W_K\right)^T}{\sqrt{d}}\right) \tag{9}$$

$$[Q_s, Q_o] = A_{score} \cdot \left([E_s, E_o]W_V\right), \tag{10}$$

where Q_s and Q_o respectively represent subject and object features, W_Q, W_K, W_V are learnable weight parameters, and d is a scaling factor.

Next, within the visual attention module and the entity attention module, we construct node-centric feature maps and edge-centric feature maps, respectively, enabling information propagation and feature fusion across these two types of graph-structured networks.

Formally, given the current hidden states of nodes and edges, denoted as H_i and $H_{(i,j)}$, we define the message for updating the i-th node as M_i, which is computed as a function of its own hidden state H_i, the hidden states of its outgoing edges $H_{(i,j)}$, and the hidden states of its incoming edges $H_{(j,i)}$. Similarly, the message for updating the edge from the i-th node to the j-th node is denoted as $M_{(i,j)}$, computed as a function of its own hidden state $H_{(i,j)}$, and the hidden states of its subject node H_i and object node H_j. More specifically, M_i and $M_{(i,j)}$ are calculated using the following two adaptive weighted message pooling functions:

$$M_i = \sum_{j:(i,j)} sigmoid\left(v_1^T[H_i, H_{(i,j)}]\right)H_{(i,j)} + \sum_{j:(j,i)} sigmoid(v_2^T[H_i, H_{(j,i)}])H_{(j,i)} \tag{11}$$

$$M_{(i,j)} = sigmoid\left(w_1^T[H_i, H_{(i,j)}]\right)H_i + sigmoid\left(w_2^T[H_j, H_{(i,j)}]\right)H_j \tag{12}$$

where[·] represents the cascade of vectors, w_1, w_2 and v_1, v_2 are learnable parameters. Next, the features are enhanced through task queries to restrict the content of the scene graph. Q_t is a task query used to calculate the attention value of task semantics and the objects with relationships in the scene graph:

$$\begin{aligned} Q_s &= [M_1 \ldots\ldots M_i \mid i \in (i,j), i = 1, \ldots, n], \\ Q_o &= [M_1 \ldots\ldots M_j \mid j \in (i,j), j = 1, \ldots, n] \end{aligned} \tag{13}$$

First, calculate the attention value between the object and the task query:

$$\alpha_i = soft\max(Q_t w_i^q [M_i w_i^k]^T) \tag{14}$$

$$[Q_s, Q_o] = \sum_{i=0}^{n} \alpha_i [Q_s, Q_o] w_i^v \tag{15}$$

Then calculate the attention value between entity relationships and task queries:

$$\alpha_{(i,j)} = soft\max(Q_t w_{(i,j)}^q [M_{(i,j)} w_{(i,j)}^k]^T) \tag{16}$$

$$[Q_{in}, Q_{out}] = \sum_{j=0}^{n} \sum_{i=0}^{n} \alpha_{(i,j)} [Q_{in}, Q_{out}] w_{(i,j)}^v \tag{17}$$

where α_i, $\alpha_{(i,j)}$ are entity attention weight and relationship attention weight, respectively. w_i^q, $w_{(i,j)}^q$, w_i^k, $w_{(i,j)}^k$, w_i^v, $w_{(i,j)}^v$ are learnable weight parameters. After multiple iterations, the final values of Q_s, Q_o, Q_{in}, Q_{out} output are used to predict object categories and relationship types.

3.3. Compositional Reasoning Execution

We have predefined the basic sub-functional modules, upon acquiring the relevant visual cues and establishing precise task decomposition, the symbolic reasoning framework systematically integrates these inputs through a structured process of combinatorial logic. This methodology employs a phased execution strategy, where specialized sub-functional modules are orchestrated to perform sequential sub-operations, ultimately synthesizing the intermediate results into a conclusive solution.

4. Experiment

To evaluate the performance of the VQA model based on the Scene Information Extraction Network (SIEN-VQA) proposed in this work, experiments were performed on both GQA and CLEVR-Human datasets and the performance of the VQA model was compared to other VQA models, an introduction to the experimental setup and result analysis is provided below.

4.1. Datasets and Setting

A. Dataset Selection

CLEVR-Humans: The dataset was written by workers on Amazon Mechanical Turk according to CLEVR images, different from CLEVR dataset, CLEVR-Hu-

mans dataset has 32164 natural language question-answer pairs which exhibit more linguistic variety than synthetic CLEVR questions, and hence be more challenging.

Table 2
Experiment result of reasoning step generation on GQA dataset.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	RSA
NS-VQA	0.984	0.971	0.959	0.946	80.2
SIEN-VQA	0.990	0.979	0.969	0.958	84.0

GQA [21]: The dataset is designed for real-world visual reasoning and combinatorial question answering, drawing inspiration from the CLEVR task and comprising 113000 images from COCO and Flickr and 22 million distinct questions, each image is annotated with a dense Scene Graph, representing the objects, attributes and relations it contains. Each question is associated with a functional program which lists the series of reasoning steps needed to be performed to arrive at the answer. Each answer is augmented with both textual and visual justifications, pointing to the relevant region within the image. GQA questions tend to involve more elements from the image compared to VQA questions, and are longer and more compositional as well. Conversely, VQA questions tend to be a bit more ambiguous and subjective, at times with no clear and conclusive answer. GQA provides more questions for each image and thus covers it more thoroughly than VQA.

The GQA dataset captures real-world visual complexity through semantically diverse scenes, whereas CLEVR-Human provides a controlled synthetic environment with explicitly annotated structural dependencies. Specifically, CLEVR-Human focuses on evaluating systematic structural reasoning capabilities through predefined compositional rules, while GQA emphasizes testing models' compositional generalization across open-domain visual-linguistic interactions. This dual evaluation paradigm leveraging both real-world and synthetic benchmarks effectively mitigates overfitting risks to domain-specific biases while enhancing cross-domain generalizability verification. Such complementary validation proves critical for developing robust visual reasoning systems capable of handling both structured logic operations and open-world semantic variations.

B. Experimental Environment Setup

The model was trained on one NVIDIA RTX A4000 GPU in a Linux Ubuntu 18.04 experimental environment, Python 3.8, CUDA 11.0, PyTorch 1.7.1.

C. Evaluation Metrics

RSA: Reasoning step accuracy, measuring whether each intermediate reasoning step is correct.

Accuracy: Accuracy of answering questions, this metric measuring whether the final answer of the model output is correct or not, directly reflects the task completion effect.

BLEU: In addition to accuracy metrics, we also used the BLEU metric to evaluate the performance of our models. BLEU measures the similarity between two sequences, and it can be divided into four common indicators: BLEU-1, BLEU-2, BLEU-3, and BLEU-4 based on n-gram, where n represents the number of consecutive elements in the sequence. We divided the generated sequence and standard sequence into multiple sub-sequence sets according to n-gram, and then investigated how many corresponding subsequences in these two sets. We used BLEU to evaluate the models' ability to select reasoning modules. BLEU is calculated as follows:

$$\text{BLEU - N} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} \omega_n \log p_n\right), \tag{18}$$

where p_n is the n-gram precision, $\omega_n = 1/\text{N}$, and BP (brevity penalty) penalizes short answers.

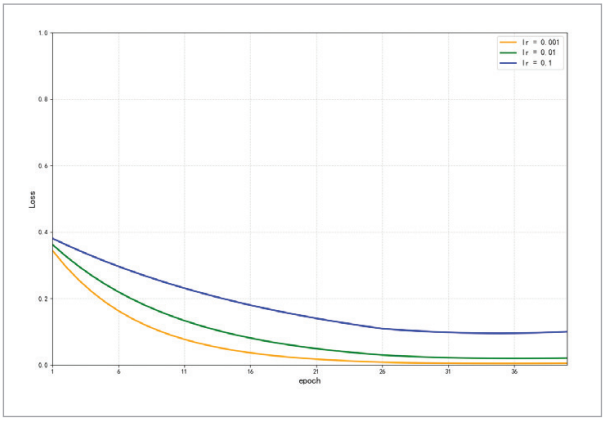
Standard metrics: GQA introduces a comprehensive evaluation framework that extends beyond conventional accuracy metrics, incorporating novel diagnostic measures to assess cross-context consistency, semantic validity, and response plausibility. This multi-dimensional metric suite enables granular behavioral analysis of reasoning models, exposing latent deficiencies in logical coherence and world knowledge grounding that traditional single-score evaluations might obscure. **Consistency** is a metric for the level of consistency in responses across different questions. **Validity** measures whether the model gives valid answers, ones that can be theoretically correct for the question. **plausibility** measures whether the model responses are reasonable in the real world or not making sense. **Distribution** measures the overall match between the true answer distribution and the model predicted distribution.

4.2. Evaluation and Analysis on GQA Dataset

Our VQA model SIEN-VQA underwent evaluation on the GQA dataset, the evaluation results were presented in Tables 2-3. In the experimental process, we set the initial learning rates to 0.1, 0.01, and 0.001, respectively, and recorded the changes in the loss function of the model during the training process as the number of epochs increased, as shown in the Figure 5. Finally, we determined the initial learning rate to be 0.001, the overall training iteration was 30 epochs, and the batch size was set to 6.

Figure 5

Visualization of hyperparameter lr.



As a first step, we evaluated the performance of the reasoning step generation on the validation set, Table 2 demonstrates that SIEN-VQA achieved an accuracy of 84.0%, which is 3.8% higher than that of the comparative model NS-VQA. SIEN-VQA outperforms the comparative model in terms of BLEU metric, indicating its ability to enhance reasoning step generation accuracy.

Table 3

Experiment result on GQA dataset.

Method	Scene Graph Type	Accuracy (%)
CRF [35]	-	72.1
LXMERT [43]	-	59.8
Lightweight [36]	Annotated	77.9
SIEN-VQA	Annotated	77.4
GraphVQA [29]	Generated	29.7
SelfGraphVQA [42]	Generated	54.0
SIEN-VQA	Generated	63.6

Table 3 presents our evaluation of the model's accuracy in answering questions on the GQA dataset after generating reasoning steps, comparing SIEN-VQA with the other VQA model based on scene graphs. In the table, '-'means there is no scene graph used. 'Annotated' indicates that the scene graph used by the model is well-annotated. 'Generated' indicates that the scene graph is generated by the model itself from the image. In addition, from the comparison in the table, it can be seen that the accuracy of the model using the generated scene map is lower than that of the model using annotated data. This indicates that the bottleneck of the current visual question answering task lies in the acquisition of visual clues, and accurate detection of scene entities and their relationships is the key to improving the accuracy of question answering. Our model has an accuracy of 77.4% with annotated data and 63.6% without annotated data, which is higher than other models that with generated scene graphs. This indicates that SIEN-VQA can obtain visual clues related to the task after parsing reasonable reasoning steps.

In Table 4, we present the evaluation results with standard metrics, and the results show that SIEN-

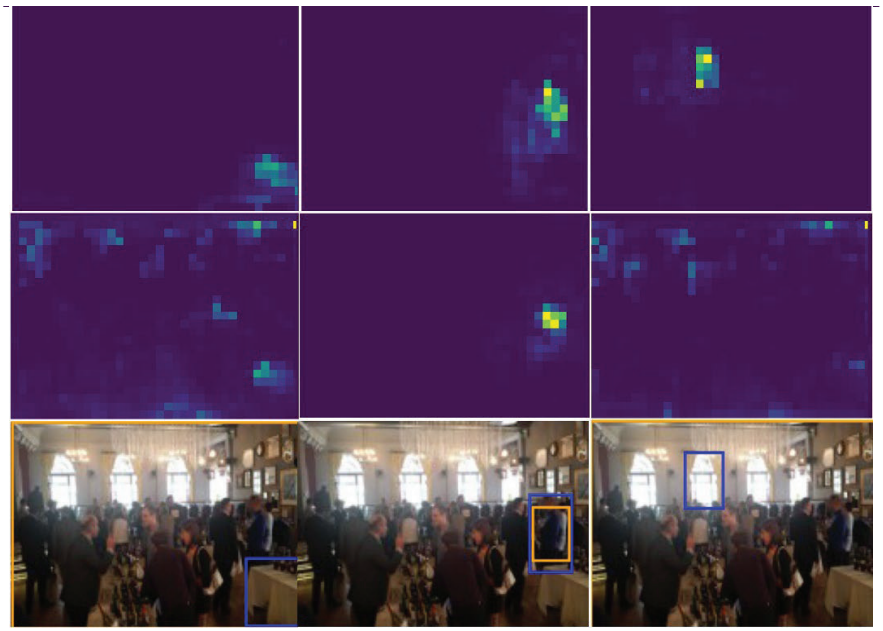
Table 4

Experiment result on GQA dataset with standard metrics.

Method	Consistency	Validity	Plausibility	Distribution	Accuracy
MAC [20]	81.6	84.5	96.2	5.3	54.1
LXMERT	89.6	84.5	96.4	5.7	59.8
NSM [22]	93.3	84.3	96.4	3.7	63.2
PVR [27]	91.4	84.8	96.5	6.0	59.8
SNMN [16]	85.1	84.8	96.4	5.1	56.1
MMN [8]	92.5	84.6	96.2	5.5	60.8
SIEN-VQA	94.4	86.0	95.5	5.2	63.6

Figure 6

Visualization results of experiment on GQA. (a) The visualization results of scene information extracting; (b) The visualization result of question answering.



(a)

Question: Is the window dark and bright? **Answer:** no

Reasoning steps:

(NS-VQA**)**: select (window) → verify (bright) → verify color (dark) → and ×

(SIEN-VQA**)**: select (window) → verify (bright) → verify tone (dark) → and √

Question: Is there any racket or bike in this picture? **Answer:** yes

Reasoning steps :

(NS-VQA**):** select (racket) → exist→ select (bike) → exist→ or √

(SIEN-VQA**)**: select (bike) → exist → select (racket) → exist→ or √

Question: Are there horses or goats? **Answer:** yes

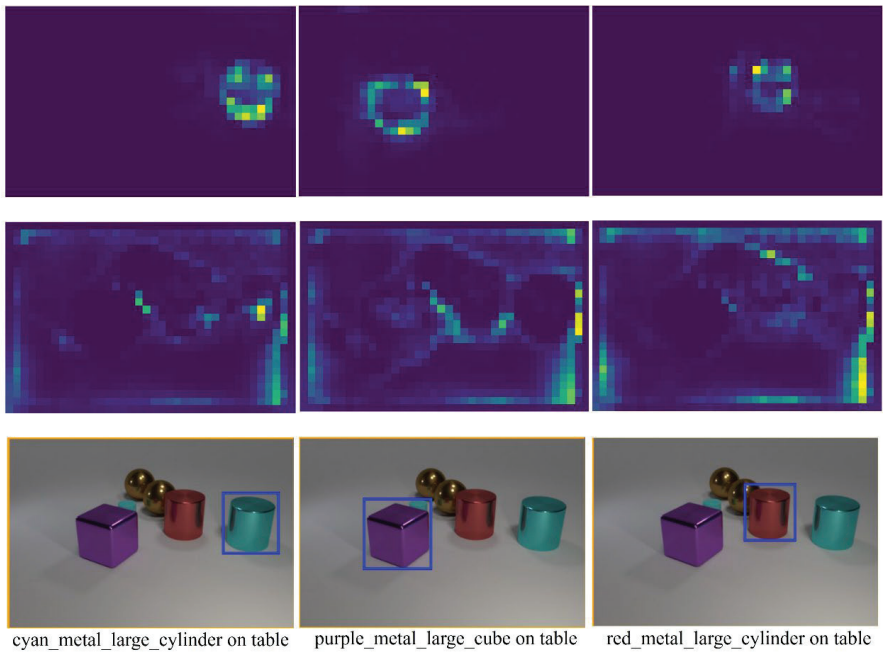
Reasoning steps :

(NS-VQA**)**: select (horse) → exist →select (goat) → exist → or √

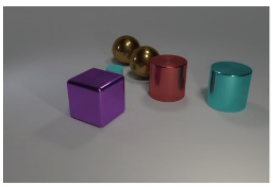
(SIEN-VQA**)**: select (horse) → exist → select (goats) → exist → or √

(b)

Figure 7
Visualization results of experiment on CLEVR-Human. (a) The visualization result of scene information extracting;(b) The visualization result of question answering



(a)

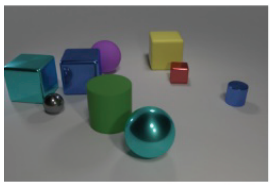


Question: There are two big metal objects at back, are they share same color?

Answer: no

Reasoning steps:

(SIEN-VQA): select (metal) → select(large)→ equal_color ×

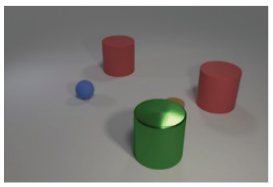


Question: How many spheres are near a cylinder?

Answer: 2

Reasoning steps :

(SIEN-VQA): select_relation (near) → find_object(cylinder) → find_subject (sphere) → count √



Question: What color is the most hidden object?

Answer: brown

Reasoning steps :

(SIEN-VQA): select_relation (hidden) → find_object( )→ query_attribute(color) √

(b)

VQA not only outperforms other comparative models in accuracy, but also performs better in consistency and validity. However, there is no significant advantage in plausibility and distribution. And we present the visualization results of the experiment in Figure 6. It can be seen that entities such as tables, people, windows, etc. have been located in the image, and the correct relationships between entities have been correctly predicted, and there is significant spatial correspondence between the heatmap's intensity peaks and the core visual entities referenced in the interrogative clause, indicating that the model has indeed achieved scene visual information delineation based on task constraints.

4.3. Evaluation and Analysis on CLEVR-Human Dataset

We also conducted experiments on CLEVR-Human. The CLEVR-Human dataset involves more flexible natural language and more complex reasoning, but the scene contents of images are simpler than that of GQA. The experiment results were shown in Table 5. The accuracy of SIEN-VQA in answering questions on CLEVR-Human is 68.5%, which is higher than NA-VQA by 1.5%. It can be seen that for simple questions the reasoning steps and scene graph generation are both effective. However, for complex reasoning problems, the completeness of scene information is also difficult to guarantee when it is unable to generate the correct reasoning steps. The visualization results are shown in Figure 7. The attention peak in the figure corresponds to the objects involved in the question stem, and based on this, the correct prediction of the relationship between objects is given.

Table 5
Experiment result on CLEVR-Human dataset.

Method	Accuary (%)
PG+EE [24]	66.0
RN [38]	57.6
MAC	50.2
RAMEN [41]	57.8
NS-VQA	67.0
SIEN-VQA	68.5

5. Conclusion

This work solved VQA task with a neural-symbolic method and proposed a VQA model based on Scene Information Extraction Network (SIEN-VQA), a Task Decomposition Network was constructed to generate reasoning steps about problems that improve task decomposition efficiency by utilizing graph structured data with the syntactic structure information of natural language, besides that, a Scene Information Extraction Network was constructed to use the generated reasoning steps as prompts, enabling the network to extract task related information, including entity categories, relationships between entities, and ignore irrelevant information. Finally, the reasoning execution components were called according to the reasoning steps to obtain the answer. The experiment results on the GQA and CLEVR-Human datasets indicate that SIEN-VQA performs well in generating reasoning steps and extracting task related information, and has higher accuracy in answering questions than the comparative model. The research in this work elaborated on the fact that reasonable reasoning steps, complete visual information, and efficient execution of reasoning are crucial elements in VQA tasks, however, the overall reasoning ability of the system is limited by the underlying reasoning execution components. The model must generate reasoning steps within the scope of the reasoning execution components, otherwise reasoning cannot be executed. Therefore, how to further improve the reasoning execution components, expand the reasoning ability of the model, and enable it to solve more complex problems is a meaningful future direction.

References

1. Agrawal, A., Lu, J., Antol, S., et al. VQA: Visual Question Answering. In Proceedings of the IEEE International Conference on Computer Vision, 2015, 2425-2433. https://doi.org/10.1109/ICCV.2015.279

2. Alper, M., Fiman, M., Averbuch-Elor, H. Is BERT Blind? Exploring the Effect of Vision-and-Language Pretraining on Visual Language Understanding. In Proceedings of the IEEE/CVF Conference on Comput-

er Vision and Pattern Recognition, 2023, 6778-6788. https://doi.org/10.1109/CVPR52729.2023.00655

3. Andreas, J., Rohrbach, M., Darrell, T., Klein, D. Learning to Compose Neural Networks for Question Answering. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, 1545-1554. https://doi.org/10.18653/v1/N16-1181
4. Andreas, J., Rohrbach, M., Darrell, T., Klein, D. Neural Module Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, 39-48. https://doi.org/10.1109/CVPR.2016.12
5. Barra, S., Bisogni, C., Marsico, M. D., Ricciardi, S. Visual Question Answering: Which Investigated Applications. Pattern Recognition Letters, 2021, 151, 325-331. https://doi.org/10.1016/j.patrec.2021.09.008
6. Ben-Younes, H., Cadene, R., Cord, M., Thome, N. MUTAN: Multimodal Tucker Fusion for Visual Question Answering. In Proceedings of the IEEE International Conference on Computer Vision, 2017, 2631-2639. https://doi.org/10.1109/ICCV.2017.285
7. Chen, L., Li, B., Shen, S., Yang, J., Li, C., Keutzer, K., Darrell, T., Liu, Z. COLA: Large Language Models Are Visual Reasoning Coordinators. In Proceedings of the International Conference on Neural Information Processing Systems, 2023, 3072, 70115-70140.
8. Chen, W., Gan, Z., Li, L., Cheng, Y., Wang, W., Liu, J. Meta Module Network for Compositional Visual Reasoning. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 2021, 655-664. https://doi.org/10.1109/WACV48630.2021.00070
9. Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P. N., Hoi, S. C. InstructBLIP: Towards General-Purpose Vision-Language Models with Instruction Tuning. arXiv Preprint, 2023, arXiv:2305.06500v2.
10. Deng, A., Cao, T., Chen, Z., Hooi, B. Words or Vision: Do Vision-Language Models Have Blind Faith in Text? arXiv Preprint, 2025, arXiv:2503.02199.
11. Eiter, T., Higuera, N., Oetsch, J., Pritz, M. A Neuro-Symbolic ASP Pipeline for Visual Question Answering. Theory and Practice of Logic Programming, 2022, 22(5), 739-754. https://doi.org/10.1017/S1471068422000229
12. Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2016, 457-468. https://doi.org/10.18653/v1/D16-1044
13. Gao, P., Jiang, Z., You, H., Lu, P., Hoi, S., Wang, X., Li, H. 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, 2019, 6639-6648. https://doi.org/10.1109/CVPR.2019.00680
14. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D. 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, 2017, 6325-6334. https://doi.org/10.1109/CVPR.2017.670
15. Hu, R. Structured Models for Vision-and-Language Reasoning. Ph.D. Dissertation, Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, California, USA, 2020.
16. Hu, R., Andreas, J., Darrell, T., Saenko, K. Explainable Neural Computation via Stack Neural Module Networks. In Proceedings of the European Conference on Computer Vision, 2018, 53-69. https://doi.org/10.1007/978-3-030-01234-2_4
17. Hu, R., Andreas, J., Rohrbach, M., Darrell, T., Saenko, K. Learning to Reason: End-to-End Module Networks for Visual Question Answering. In Proceedings of the IEEE International Conference on Computer Vision, 2017, 804-813. https://doi.org/10.1109/ICCV.2017.93
18. Huang, L., Kulkarni, K., Jha, A., Lohit, S., Jayasuriya, S., Turaga, P. CS-VQA: Visual Question Answering with Comprehensively Sensed Images. arXiv Preprint, 2018, arXiv:1806.03379. https://doi.org/10.1109/ICIP.2018.8451445
19. Huang, T., Yang, Y., Yang, X. A Survey of Deep Learning-Based Visual Question Answering. Journal of Central South University, 2021, 28(3), 728-746. https://doi.org/10.1007/s11771-021-4641-x
20. Hudson, D. A., Manning, C. D. Compositional Attention Networks for Machine Reasoning. In Proceedings of the International Conference on Learning Representations, 2018, 1-20.
21. Hudson, D. A., Manning, C. D. 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, 2019, 6693-6702. https://doi.org/10.1109/CVPR.2019.00686
22. Hudson, D. A., Manning, C. D. Learning by Abstraction: The Neural State Machine. In Proceedings of the Advances in Neural Information Processing Systems, 2019, 5903-5916.

23. Jabri, A., Joulin, A., Van Der Maaten, L. Revisiting Visual Question Answering Baselines. In Proceedings of the European Conference on Computer Vision, 2016, 727-739. https://doi.org/10.1007/978-3-319-46484-8_44

24. Johnson, J., Hariharan, B., Van Der Maaten, L., Hoffman, J., Li, F., Zitnick, C. L., Girshick, R. Inferring and Executing Programs for Visual Reasoning. In Proceedings of the IEEE International Conference on Computer Vision, 2017, 3008-3017. https://doi.org/10.1109/ICCV.2017.325

25. Johnson, J., Hariharan, B., Van Der Maaten, L., Li, F., Zitnick, C. L., Girshick, R. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, 1988-1997. https://doi.org/10.1109/CVPR.2017.215

26. Li, F., Krishna, R. Searching for Computer Vision North Stars. Daedalus, 2022, 151(2), 85-99. https://doi.org/10.1162/daed_a_01902

27. Li, G., Wang, X., Zhu, W. Perceptual Visual Reasoning with Knowledge Propagation. In Proceedings of the ACM International Conference on Multimedia, 2019, 530-538. https://doi.org/10.1145/3343031.3350922

28. Li, J., Selvaraju, R. R., Gotmare, A., Joty, S., Xiong, C., Hoi, S. C. H. Align Before Fuse: Vision and Language Representation Learning with Momentum Distillation. In Proceedings of the International Conference on Neural Information Processing Systems, 2021, 9694-9705.

29. Liang, W., Jiang, Y., Liu, Z. GraphVQA: Language-Guided Graph Neural Networks for Graph-Based Visual Question Answering. In Proceedings of the Workshop on Multimodal Artificial Intelligence, 2021, 79-86. https://doi.org/10.18653/v1/2021.maiworkshop-1.12

30. Lin, X., Parikh, D. Leveraging Visual Question Answering for Image-Caption Ranking. In Proceedings of the European Conference on Computer Vision, 2016, 261-277. https://doi.org/10.1007/978-3-319-46475-6_17

31. Lu, J., Yang, J., Batra, D., Parikh, D. Hierarchical Question-Image Co-Attention for Visual Question Answering. In Proceedings of the International Conference on Neural Information Processing Systems, 2016, 289-297.

32. Malinowski, M., Fritz, M. A Multi-World Approach to Question Answering About Real-World Scenes Based on Uncertain Input. In Proceedings of the International Conference on Neural Information Processing Systems, 2014, 1682-1690.

33. Malinowski, M., Rohrbach, M., Fritz, M. Ask Your Neurons: A Neural-Based Approach to Answering Questions About Images. In Proceedings of the IEEE International Conference on Computer Vision, 2015, 1-9. https://doi.org/10.1109/ICCV.2015.9

34. Manmadhan, S., Kovoor, B. C. Visual Question Answering: A State-of-the-Art Review. Artificial Intelligence Review, 2020, 53(8), 5705-5745. https://doi.org/10.1007/s10462-020-09832-7

35. Nguyen, B. X., Do, T., Tran, H., Tjiputra, E., Tran, Q. D., Nguyen, A. Coarse-to-Fine Reasoning for Visual Question Answering. In Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition, 2022, 4558-4566. https://doi.org/10.1109/CVPRW56347.2022.00502

36. Nuthalapati, S. V., Chandradevan, R., Giunchiglia, E., Li, B., Kayser, M., Lukasiewicz, T., Yang, C. Lightweight Visual Question Answering Using Scene Graphs. In Proceedings of the ACM International Conference on Information and Knowledge Management, 2021, 3353-3357. https://doi.org/10.1145/3459637.3482218

37. Ren, M., Kiros, R., Zemel, R. Exploring Models and Data for Image Question Answering. In Proceedings of the International Conference on Neural Information Processing Systems, 2015, 2953-2961.

38. Santoro, A., Raposo, D., Barrett, D. G., Malinowski, M., Pascanu, R., Battaglia, P., Lillicrap, T. A Simple Neural Network Module for Relational Reasoning. In Proceedings of the Advances in Neural Information Processing Systems, 2017, 4967-4976.

39. Schulze Buschoff, L. M., Akata, E., Bethge, M., Schulz, E. Visual Cognition in Multimodal Large Language Models. Nature Machine Intelligence, 2025, 7, 96-106. https://doi.org/10.1038/s42256-024-00963-y

40. Shih, K., Singh, S., Hoiem, D. Where to Look: Focus Regions for Visual Question Answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, 4613-4621. https://doi.org/10.1109/CVPR.2016.499

41. Shrestha, R., Kafle, K., Kanan, C. Answer Them All! Toward Universal Visual Question Answering Models. In Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition, 2019, 10472-10481. https://doi.org/10.1109/CVPR.2019.01072

42. Souza, B., Aasan, M., Pedrini, H., Ramírez Rivera, A. SelfGraphVQA: A Self-Supervised Graph Neural Network for Scene-Based Question Answering. 2023,

arXiv:2310.01842v1. https://doi.org/10.1109/ICCVW60793.2023.00499

43. Tan, H., Bansal, M. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. 2019, arXiv:1908.07490. https://doi.org/10.18653/v1/D19-1514
44. Teney, D., Hengel, A. Visual Question Answering as a Meta Learning Task. In Proceedings of the European Conference on Computer Vision, 2018, 229-245. https://doi.org/10.1007/978-3-030-01267-0_14
45. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V., Zhou, D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. 2022, arXiv:2201.11903v6.
46. Yamada, M., D'Amario, V., Takemoto, K., Boix, X., Sasaki, T. Transformer Module Networks for Systematic Generalization in Visual Question Answering. 2023, arXiv:2201.11316. https://doi.org/10.1109/TPAMI.2024.3438887
47. Yi, K., Wu, J., Gan, C., Torralba, A., Kohli, P., Tenenbaum, J. B. Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding. In Proceedings of the International Conference on Neural Information Processing Systems, 2019, 1039-1050.
48. Yu, D., Fu, J., Mei, T., Rui, Y. Multi-Level Attention Networks for Visual Question Answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, 4187-4195. https://doi.org/10.1109/CVPR.2017.446
49. Yu, D., Gao, X., Xiong, H. Structured Semantic Representation for Visual Question Answering. In Proceedings of the IEEE International Conference on Image Processing, 2018, 2286-2290. https://doi.org/10.1109/ICIP.2018.8451516
50. Yu, Z., Yu, J., Xiang, C., Fan, J., Tao, D. Beyond Bilinear: Generalized Multimodal Factorized High-Order Pooling for Visual Question Answering. IEEE Transactions on Neural Networks and Learning Systems, 2017, 29(12), 5947-5959. https://doi.org/10.1109/TNNLS.2018.2817340
51. Zhang, J., Huang, J., Jin, S., Lu, S. Vision-Language Models for Vision Tasks: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(8), 5625-5644. https://doi.org/10.1109/TPAMI.2024.3369699
52. Zhang, P., Goyal, Y., Summers-Stay, D., Batra, D., Parikh, D. Yin and Yang: Balancing and Answering Binary Visual Questions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, 5014-5022. https://doi.org/10.1109/CVPR.2016.542
53. Zhou, B., Tian, Y., Sukhbaatar, S., Szlam, A., Fergus, R. Simple Baseline for Visual Question Answering. 2015, arXiv:1512.02167.
54. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. 2022, arXiv:2304.10592.
55. Zhu, Y., Groth, O., Bernstein, M., Li, F. Visual7W: Grounded Question Answering in Images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, 4995-5004. https://doi.org/10.1109/CVPR.2016.540