

# Knowledge-Augmented Visual Question Answering With Natural Language Explanation

Jiayuan Xie<sup>1</sup>, Yi Cai<sup>2</sup>, *Member, IEEE*, Jiali Chen, Ruohang Xu, Jiexin Wang<sup>3</sup>, *Member, IEEE*, and Qing Li<sup>4</sup>, *Fellow, IEEE*

**Abstract**—Visual question answering with natural language explanation (VQA-NLE) is a challenging task that requires models to not only generate accurate answers but also to provide explanations that justify the relevant decision-making processes. This task is accomplished by generating natural language sentences based on the given question-image pair. However, existing methods often struggle to ensure consistency between the answers and explanations due to their disregard of the crucial interactions between these factors. Moreover, existing methods overlook the potential benefits of incorporating additional knowledge, which hinders their ability to effectively bridge the semantic gap between questions and images, leading to less accurate explanations. In this paper, we present a novel approach denoted the knowledge-based iterative consensus VQA-NLE (KICNLE) model to address these limitations. To maintain consistency, our model incorporates an iterative consensus generator that adopts a multi-iteration generative method, enabling multiple iterations of the answer and explanation in each generation. In each iteration, the current answer is utilized to generate an explanation, which in turn guides the generation of a new answer. Additionally, a knowledge retrieval module is introduced to provide potentially valid candidate knowledge, guide the generation process, effectively bridge the gap between questions and images, and enable the production of high-quality answer-explanation pairs. Extensive experiments conducted on three different datasets demonstrate the superiority of our proposed KICNLE model over competing state-of-the-art approaches. Our code is available at <https://github.com/Gary-code/KICNLE>.

**Index Terms**—Visual question answering, natural language explanation, multimodal.

Manuscript received 17 June 2023; revised 7 November 2023 and 24 January 2024; accepted 25 February 2024. Date of publication 28 March 2024; date of current version 5 April 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62076100; in part by the Fundamental Research Funds for the Central Universities, South China University of Technology (SCUT), under Grant x2rjD2230080; in part by the Science and Technology Planning Project of Guangdong Province under Grant 2020B0101100002; in part by Guangdong Provincial Fund for Basic and Applied Basic Research—Regional Joint Fund Project (Key Project) under Grant 23201910250000318 and Grant 308155351064; in part by the Chinese Association for Artificial Intelligence (CAAI)-Huawei MindSpore Open Fund; in part by the China Computer Federation (CCF)-Zhipu AI Large Model Fund; and in part by the Hong Kong Research Grants Council through the Theme-Based Research Scheme under Project T22-501/23-R (Sub-Project: P0049913). The associate editor coordinating the review of this manuscript and approving it for publication was Mr. Chuang Gan. (*Corresponding author: Jiexin Wang.*)

Jiayuan Xie and Qing Li are with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong, SAR, China.

Yi Cai, Jiali Chen, Ruohang Xu, and Jiexin Wang are with the School of Software Engineering and the Key Laboratory of Big Data and Intelligent Robot, Ministry of Education, South China University of Technology, Guangzhou 510006, China (e-mail: jiexinwang@scut.edu.cn).

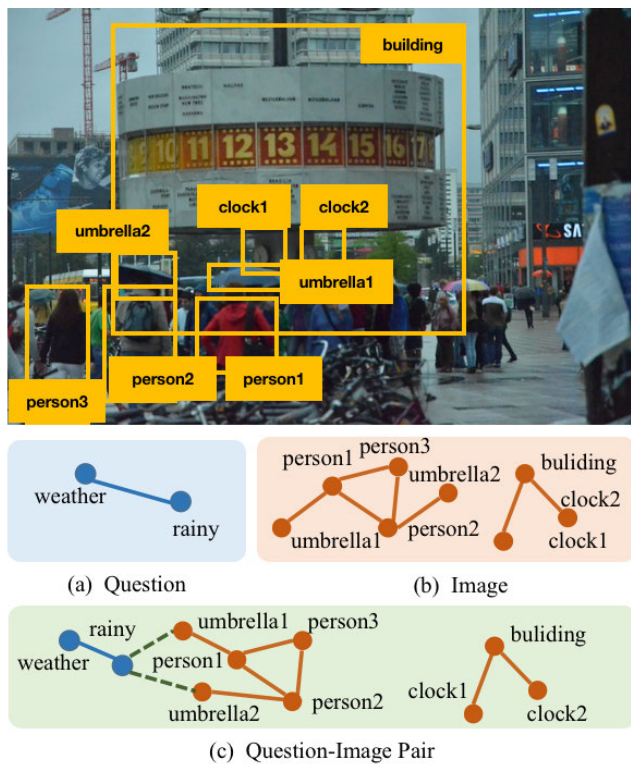
Digital Object Identifier 10.1109/TIP.2024.3379900

## I. INTRODUCTION

VISUAL question answering (VQA) systems infer the relationships between given question-image pairs and generate answers based on an answering decision-making process that is determined by these relationships [1], [2]. However, most existing studies on VQA focus solely on answer generation and neglect the generation of explanations that could reveal the underlying decision-making process, which leads to a general depiction of the models as black boxes [3]. Consequently, VQA with natural language explanation (VQA-NLE) has emerged as a significant research direction within the VQA task and has attracted increasing attention from researchers in recent years [4], [5]. A recent study [3] also strongly emphasized the significance of elucidating the decision-making process in VQA models for various reasons, including establishing trust, ensuring accountability, enhancing the understanding of model biases, and promoting correctness. Thus, the task of VQA-NLE involves generating answers and elucidating the decision-making process of VQA models using human-friendly natural language sentences.

Compared with traditional vision-language generation tasks like image captioning [6], [7], [8], [9], VQA-NLE has two distinctive challenges, (i) **Consistency**: the generated explanations should exhibit robust coherence with the answers rather than just being related to the given question-image pairs. As illustrated in Figure 1, when the model predicts the answer to be “Yes” in response to the question “Does the weather appear rainy?”, the explanation should precisely describe what is consistent with “rainy weather”. Conversely, when the predicted answer is “No”, the explanation should reflect the opposing characteristics. (ii) **Correlation**: the generated explanations should be grounded in inferred relationships that establish a correlation between question-image pairs. As also shown in Figure 1, the model needs to bridge the semantic gap between the textual mention of “rainy weather” in the question and the visual cue of “umbrellas” in the image, to generate an explanation related to both components.

Current studies on the VQA-NLE task mainly jointly encode image vision information and question language information to generate answers and explanations either at the same time or in a step-by-step manner. However, the lack of interaction in the generation process, such as not considering the explanation during the answer generation process, poses challenges in ensuring the consistency between the generated explanation and the answer. Moreover, these studies also are limited in



**Q:** Does the weather appear rainy? **A:** Yes

**VQA-X Explanation:** Lots of people are using umbrellas.

**Our Explanation (VQA-X-KB):** Lots of people are using umbrellas, and umbrellas are used for rainy days.

**(Knowledge: an [[umbrella]] UsedFor [[rainy days]])**

Fig. 1. An example from the VQA-X dataset, consisting of a question-image pair, a correct answer, and its corresponding explanation. The explanation needs to bridge the gap between the question and the image, i.e., connecting the image object “umbrella” with the “rainy” in the question. Furthermore, we integrate knowledge into the explanations of the existing VQA-X dataset to make them more aligned with real-world reasoning, resulting in the creation of our VQA-X-KB dataset.

properly bridging the gap between images and questions, especially when encountering implicit relationships like that between “rainy” and “umbrella”. For the first challenge, a pedagogical study [10] demonstrated that humans typically adopt a multi-iteration feedback method to enhance consistency between answers and explanations. In each iteration, the current answer is used to generate an explanation, and subsequently, the generated explanation is employed to generate a new answer. For the second challenge, establishing accurate correlations between question-image pairs typically requires humans to integrate their additional knowledge. For example, in the case of Figure 1, the knowledge that “an umbrella is used for rainy days” is crucial for establishing a proper correlation.

In this paper, we present Knowledge-augmented Iterative Consensus VQA-NLE (KICNLE), a model designed to address the challenges of consistency and correlation. The model comprises three essential modules: an original information extractor, a knowledge retrieval module, and an iterative consensus generator. The original information extractor utilizes

ViT [11] to extract patch-level image features, and employs the Oscar [12] to obtain multimodal features of aligned images and questions. To address the challenge of consistency, in the aforementioned pedagogical research, the iterative consensus generator incorporates an interactive generation mechanism inspired by the multi-iteration feedback method [10]. Specifically, this mechanism involves two steps: (i) a rough-answer GPT module designed to generate an initial rough answer and (ii) an explanation GPT module utilized to generate an explanation based on the rough answer, which is subsequently used by an answer GPT module to produce a new answer. Through multiple iterations of the second step, the generated answers and explanations exhibit a high level of consistency due to their mutual consideration of each other in the generation process. To address the challenge of correlation, a knowledge retrieval module is introduced to bridge the gap between questions and images; this module provides each GPT module with potentially valid candidate knowledge for guidance during generation. Finally, our model integrates additional knowledge to accurately determine question-image pair relationships, resulting in consistently high-quality answers and explanations.

Our main contributions are summarized as follows:

- To the best of our knowledge, this is the first work to explore the integration of additional knowledge in the VQA-NLE task. By incorporating such supplementary knowledge, our proposed KICNLE model effectively bridges the gap between images and questions, enables the generation of more accurate answers and provides more informed explanations.
- Our model introduces an iterative consensus generator that adopts an interactive generative mechanism inspired by pedagogical research, which enforces consistency between the generated answers and explanations through a multi-iteration feedback method.
- We construct a new dataset called VQA-X-KB by extending the existing VQA-X dataset. The new dataset integrates additional knowledge into the original explanations, making these explanations more reasonable. For example, in Figure 1, the knowledge that “an umbrella used for rainy days” is added to the explanation.
- Extensive experiments on three datasets (VQA-X [13], VQA-X-KB, and A-OKVQA [14]) show that our proposed KICNLE model can effectively bridge the gap and achieve new state-of-the-art performance compared to the previous VQA and NLE models.

## II. RELATED WORK

The visual question answering with natural language explanation (VQA-NLE) task seeks to produce a reliable answer and its corresponding explanation; this task consists of two subtasks, i.e., visual question answering [12], [15], [16], [17], [18], [19], [20], [21] and explanation generation [3], [4], [13], [22], [23], [24], [25].

### A. Visual Question Answering

The existing methods for VQA can be broadly categorized into two main approaches: the basic encoder-decoder

framework [15], [16], [17], [18] and the large-scale pretrained model [12], [19], [20], [21]. In the basic encoder-decoder framework, visual features of the input image and textual features of the query text are first extracted and subsequently fused to generate answers. Attention mechanisms have been widely exploited to improve the performance of these models. For example, Xu and Saenko [15] apply a spatial memory network [26] to align words in a question with image patches and select relevant visual evidence to predict the answer based on the alignment. Yu et al. [16] use self-attention and guided-attention modules to fuse the intramodal and inter-modal information between a given question and an image. Furthermore, the external knowledge can be leveraged as helpful information for VQA. Wu and Mooney [17] introduce an entity-focused retrieval (EnFoRe) model, which is able to recognize question-relevant entities and retrieve specific knowledge about those entities from Wikipedia to predict the answer. Guo et al. [18] propose a unified retriever-reader framework that combines the implicit knowledge and the explicit knowledge towards knowledge-based VQA. Specifically, the designed pseudo labels are utilized for effective explicit knowledge retrieval, while implicit knowledge is obtained from a vision-language pretrained model. Moreover, some studies focus on the neurosymbolic visual question answering [27], [28], [29], [30], which recovers a structural scene representation from an image and then extracts a program sequence from the question. Specifically, Gan et al. [30] collect visual questions and segmentation answers, which link the instance segmentation to questions and answers in the VQA dataset. Yi et al. [29] propose a model to recognize the objects in the scene and then a symbolic program executor is used for reasoning and answering questions. Mao et al. [28] design a framework to look at images and analyze associated questions, without supervision object labels. Barbiero et al. [27] propose a deep concept reasoner to build syntactic rule structures using concepts and execute these rules on meaningful concepts for an interpretable prediction. On the other hand, large-scale pretrained vision-language models have gained increasing attention in recent years due to their potential applications in various multimodal downstream tasks, including visual question answering. For instance, Chen et al. [19] propose the UNITER model learned through four image-text datasets to obtain joint embeddings of images and texts for VQA. Li et al. [21] design a unified Transformer-based pretrained model that aligns the textual and visual information with the noisy web data bootstrapping strategy.

### B. Explanation Generation in VQA

In the VQA process, the generation of explanations is critical in realizing an explicit reasoning process of answers, thereby enhancing the overall reliability of the model. Ayyubi et al. [22] introduce the explanation generation task of generating rationales of VQA as a measure of the model's comprehensive understanding. This task requires the model to rationalize the predicted answer based on the question and image. Different approaches have been proposed to address

this task. Park et al. [13] utilize the MCB [31] model for answer prediction and then use an LSTM-based [32] model for explanation generation. Wu and Mooney [23] utilize an upgraded Up-Down VQA [33] model to generate answers and use an LSTM-based model to generate explanations based on the gradient-based visual attention features, which is consistent with the predicted answers. Marasovic et al. [24] develop an integrated model to generate free-text rationales by combining pretrained language models with object recognition, grounded visual semantic frames, and visual common sense graphs. Kayser et al. [25] adopt the pretrained vision-language model UNITER [19] to jointly encode a given question and image, and subsequently apply the GPT-2-based [34] model to explain the results. However, the generated answer and explanation are not causally related since they are generated by two relatively independent models. To alleviate this problem, Sammani et al. [3] propose the NLX-GPT, which utilizes a language model (i.e., GPT-2) to generate the answer and explanation simultaneously. Yang et al. [4] introduce a chunk-aware semantic alignment module to first capture the chunk-level semantics based on questions and images. Next, the token-level and chunk-level multimodal representations are used for answer inference. Finally, lexical constraints associated with words or chunks are incorporated to explain the results.

To the best of our knowledge, none of the previous works have integrated knowledge to determine the relationships between question-image pairs for generating explanations. This limitation is present even though integrating knowledge can enhance the reasoning process and contribute to the generation of more informative, reliable and high-quality explanations.

## III. METHODOLOGY

Given an image  $I$  and a question  $Q = [q_1, \dots, q_L]$  associated with the image, the goal of the VQA-NLE task is to generate an answer  $A = [a_1, \dots, a_{L_a}]$  and an explanation  $E = [e_1, \dots, e_{L_e}]$  that accurately reflect the decision-making process. Here,  $L$ ,  $L_a$  and  $L_e$  denote the length of the given question, the generated answer and the explanation, respectively, and these lengths are not fixed constants. Compared with traditional vision-language generation tasks, VQA-NLE requires bridging the gap between question-image pairs to generate coherent answers and explanations. Leveraging this insight, we propose the knowledge-augmented iterative consensus VQA-NLE (KICNLE) model, and the overall framework of our model is depicted in Figure 2. Our model consists of three modules: (i) The original information extractor aims to extract the semantic features contained in the original information (i.e., the given images and questions). (ii) The knowledge retrieval module is applied to retrieve potentially effective candidate knowledge that helps establish connections between questions and images. (iii) An iterative consensus generator generates the answer and the explanation, and it corrects any inconsistencies that arise between them. The details of each component of the KICNLE model are presented in the following sections.



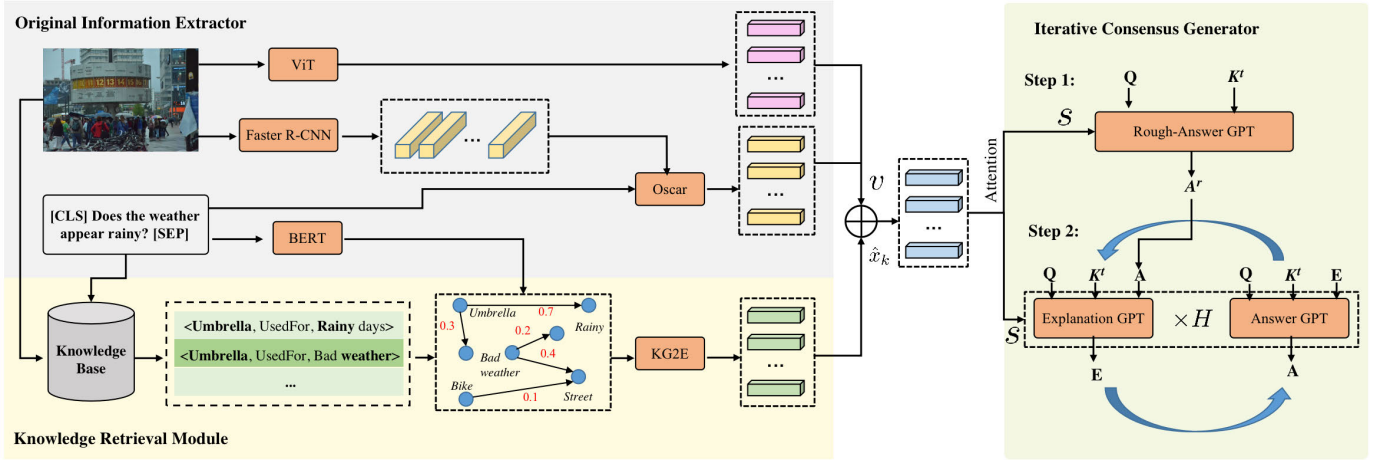


Fig. 2. Overview of our framework. It contains three components: (i) the original information extractor with a grey color background, (ii) the knowledge retrieval module with a yellow color background, and (iii) the iterative consensus generator with a green color background.

### A. Original Information Extractor

The original information in the VQA task consists of two parts, i.e., visual images and textual questions. For a given image  $I$ , we first apply the CLIP vision encoder [11] to extract the image features, which provide complete and detailed image information. Specifically, we divide the given image with a resolution of  $H \times W$  into  $P$  patches. These patches are then passed through the CLIP vision encoder, resulting in patch-level features denoted as  $\{v_i^s\}_{i=1}^P$ , where  $\{v_i^s\} \in \mathbb{R}^d$  and  $d$  represent the feature dimensions. Furthermore, because the content involved in a given question in the VQA-NLE task is usually related to specific objects in the image, we then employ an off-the-shelf pretrained Faster R-CNN [35] to extract the object-level visual feature  $\{v_i^o\}_{i=1}^N$  of the image, where  $\{v_i^o\} \in \mathbb{R}^{d'}$  and  $N$  is the number of objects.

Because images and questions are represented with different modalities, it is imperative to encode the question text while aligning this texts with appropriate image features. To achieve this integration, we leverage a pretrained vision-language model called Oscar [12], which enables us to obtain fine-grained multimodal representations that capture the alignment between images and questions. For a given question  $Q$ , we first utilize a tokenizer to tokenize the question, resulting in a sequence of question tokens. Special tokens [CLS] and [SEP] are then added to denote the start and end of the question sequence, respectively. Subsequently, the object-level visual features  $v^o$  and the question sequence are jointly processed by multilayer transformers within the Oscar model, which enables the model to capture intricate relationships and interactions among the vision-language features. The output of the Oscar model is denoted as  $\{v_i^r\}_{i=1}^{L+N}$ , where  $v_i^r \in \mathbb{R}^d$  and  $L + N$  correspond to the sum of the question length and the number of objects, respectively.

Based on these representations, we concatenate the dual features (i.e., the output of Oscar, denoted as  $v^r$ , and the patch-level features, denoted as  $v^s$ ) to obtain the joint visual-language features  $v$ ,

$$v = [v^r \oplus v^s], \quad (1)$$

where  $v \in \mathbb{R}^d$  and  $\oplus$  denotes the concatenation operation.

### B. Knowledge Retrieval Module

To bridge the gap between images and questions, our approach begins by addressing the semantic alignment of multimodal features via the original information extractor module. However, relying solely on the semantic information of words and images may be insufficient, necessitating the incorporation of additional knowledge. For example, as depicted in Figure 1, connecting concepts such as “umbrella” and “rainy day” requires the inclusion of additional knowledge, such as “Umbrella Used for Rainy days”. Thus, the knowledge retrieval module (KRM) introduced in this subsection bridges the semantic gap and captures accurate relationships between connecting concepts through the retrieved target knowledge.

The KRM module first retrieves the candidate knowledge triplets by treating the objects in the image as the head entities of the external graph knowledge base (i.e., ConceptNet [36]). Specifically, ConceptNet is an available semantic network, designed to help computers understand the meanings of words that people use. ConceptNet originated from the open-mind common sense of crowdsourcing projects and contains more than 10 million entities involving various concepts, things, places, relationships and events. For example, in Figure 2, the KRM retrieves a knowledge triplet  $\langle \text{Umbrella}, \text{UsedFor}, \text{Rainy\_Day} \rangle$  where “Umbrella” serves as the head entity, as it corresponds to the image object “umbrella”. Considering that there is a tremendous amount of knowledge triplet in the knowledge base, we employ the following steps to obtain knowledge representation and identify the most appropriate target knowledge. We first concentrate on the top  $M$  knowledge triplets of each image object. This approach is based on the observation that in VQA, certain common knowledge tends to provide more useful information, while other knowledge may be less relevant or even unhelpful. These top  $M$  triplets are selected based on their respective “weight” parameter within ConceptNet, where the “weight” value indicates the co-occurrence probability of encountering objects of the triplet together. Considering the potential variations in the contributions of the top  $M$  triplets in bridging the image and the question, we then employ a question-aware encoder to evaluate the relationship between a given question and the triplets.

First, we use a pretrained BERT model to encode the given question  $Q$ , because this approach better calculates semantic similarity than do traditional term-based retrieval algorithms (e.g., BM25 [37] and TF-IDF [38]),

$$x_q = \text{BERT}(Q), \quad (2)$$

where  $x_q \in \mathbb{R}^d$  represents the representation of question  $Q$ .

Then, to obtain the features of the triples, we consider not only the relationship information between the head and tail entities but also the semantic information of each entity in the triple. To capture the relationship information between the head and tail entities, we follow an approach similar to that of Xie et al. [39] and utilize the KG2E [40] word embedding lookup table  $e^g(\cdot)$ . This allows us to obtain embeddings of each tail entity  $e_{i,j}$  in the knowledge triplet,

$$h_{i,j}^{ke} = e^g(e_{i,j}). \quad (3)$$

Here,  $h_{i,j}^{ke} \in \mathbb{R}^d$  denotes the representation of the  $j$ -th entity related to the  $i$ -th object of the image.

For the semantic information of the entities, we employ the BERT model to capture the semantic representations of each tail entity  $e_{i,j}$  using the [CLS] token, represented as  $h_{i,j}^w$ . Subsequently, we concatenate this representation with the KG2E embedding  $h_{i,j}^{ke}$  to obtain the final representations of the knowledge triplet  $h_{i,j}^k$ ,

$$h_{i,j}^w = \text{BERT}(e_{i,j}), \quad (4)$$

$$h_{i,j}^k = [h_{i,j}^w \oplus h_{i,j}^{ke}]. \quad (5)$$

Next, for the tail entities present in the top  $M$  candidate knowledge triplets corresponding to the  $i$ -th object, the question-aware encoder employs a cross-attention mechanism to calculate normalized weights for each tail entity. The corresponding calculation formula is expressed as,

$$s_{i,j} = \frac{\exp(W_s(h_{i,j}^k + x_q) + b_s)}{\sum_{j=0}^M \exp(W_s(h_{i,j}^k + x_q) + b_s)}, \quad (6)$$

where  $s_{i,j}$  represents the weight of the  $j$ -th knowledge triplet of the  $i$ -th object, and  $W_s$  and  $b_s$  are learnable parameters. Therefore, the knowledge representation of the  $i$ -th object can be calculated by determining the weighted sum of  $M$  triplet representations,

$$\hat{x}_i^k = \sum_{j=0}^M s_{i,j} \cdot h_{i,j}^k. \quad (7)$$

Finally, we determine the most appropriate target knowledge based on two situations. In the first situation, when the tail entity in a candidate triplet appears in the question, we utilize a word matching method to retrieve this triplet and directly use it as target knowledge  $k^t$ . In the second situation, the knowledge triplet with the highest weight is considered the target knowledge  $k^t$ . This triplet serves as the explicit textual prompt with which the decoder generates the answer and the relevant explanation.

### C. Iterative Consensus Generator

In the VQA-NLE task, the generated answers and explanations must be consistent. Inspired by the multi-iteration generation strategy that is used in human cognition [10], we design the iterative consensus generator module. This module facilitates communication between the answer and explanation generation processes, enabling them to consistently reach a consensus. The details of this generator structure are illustrated in the right half of Figure 2. The overall generation strategy is divided into the following two steps.

*Step 1 (Rough Answer Generation):* Generate a rough answer based on the given information and the target knowledge.

*Step 2 (Iterative Answer and Explanation Generation):* Use the rough answer from the first step to generate a rough explanation. Subsequently, we iteratively refine both the answers and explanations to achieve improved consistency.

To provide a clearer description of the module, we define  $A_j$  and  $E_j$  as the generated answer and explanation, respectively, in the  $j$ -th iteration, while  $A^r$  represents the generated rough answer. The generator module can therefore be concisely defined as follows:

$$A^r = \operatorname{argmax}_A P(A | I, Q, \mathcal{K}), \quad (8)$$

$$E_j = \begin{cases} \operatorname{argmax}_E P(E | I, Q, \mathcal{K}, A_{j-1}), & j > 1 \\ \operatorname{argmax}_E P(E | I, Q, \mathcal{K}, A^r), & j = 1, \end{cases} \quad (9)$$

$$A_j = \operatorname{argmax}_A P(A | I, Q, \mathcal{K}, E_j), \quad (10)$$

where  $\mathcal{K}$  contains the knowledge triplet representation  $\hat{x}_i^k$  obtained from Equation (7) and the target knowledge  $k^t$ .

Following the study of Sammani et al. [3], we adopt the Distilled GPT-2 [41] as the pretrained language model to generate the answer and explanation. Moreover, we concatenate the knowledge representation  $\hat{x}_k$  obtained from the knowledge retrieval module and the multimodal feature  $v$  obtained from the original information extractor. These concatenated representations serve as the keys and values in the standard multihead attention sublayer of Distilled GPT-2 in each generation step:

$$s = [v \oplus \hat{x}_k], \quad (11)$$

where  $\oplus$  denotes the concatenation operation.

*1) Rough Answer Generation:* To generate a rough answer, we introduce a Distilled GPT-2-based model [34] called the rough-answer GPT, where the multimodal feature with knowledge  $s$  is utilized as the encoder hidden state. In the decoder step  $t$ , the decoder module utilizes the question sequence  $Q$ , the target knowledge  $k^t$ , the last generated word embedding  $\hat{a}_{t-1}^r$  of the rough answer, and the multimodal feature with knowledge  $s$  to generate the current hidden state,

$$h_t^r = \text{Dec}(\hat{a}_{t-1}^r, s, Q, k^t). \quad (12)$$

Following previous studies [42], we develop a fully connected layer over  $h_t^r$ , learn from this layer, and then utilize softmax activation to obtain the word probability distribution:

$$p(\hat{a}_t^r) = \text{softmax}(W_{ar} h_t^r + b_{ar}), \quad (13)$$

where  $W_{ar}$  and  $b_{ar}$  are the parameters to be learned.

Considering that the accuracy of the rough answer is also important for subsequent iterative generation, we use the binary cross-entropy objective to calculate the word-level loss  $\mathcal{L}_r$  between the ground truth answer distribution  $p(a_t)$  and the word probability distribution of the rough answer,

$$\mathcal{L}_r = - \sum_{t=1}^L p(a_t) \log p(\hat{a}_t^r) + (1 - p(a_t)) \log (1 - p(\hat{a}_t^r)). \quad (14)$$

2) *Answer-Explanation Iterative Generation*: Given the multimodal feature with knowledge  $s$  and the rough answer  $A^r$  obtained in the first step, our objective is to ensure consistency between the final answer and explanation. For this purpose, we employ two Distilled GPT models, named the Answer GPT and the Explanation GPT, as the decoders for iterative generation of the answer and explanation. Both of these tools include the same structure and shared parameters, except for the input text.

At each iteration  $j$ , we refine the answer and the explanation based on the results from the previous iteration. This strategy maximizes the agreement between the generators and enables the generators to communicate with each other through the results they generate. Specifically, we use  $\hat{a}_{j,t}$  and  $\hat{e}_{j,t}$  to denote the generated answer and explanation, respectively, in the decoder step  $t$  of the iteration  $j$ . Like in rough answer generation, in iterative generation, the explanation and answer can be described as follows:

$$h_{j,t}^e = \begin{cases} \text{Dec}(\hat{e}_{j,t-1}, s, Q, k^t, A^r) & j = 1 \\ \text{Dec}(\hat{e}_{j,t-1}, s, Q, k^t, A_{j-1}) & j > 1, \end{cases} \quad (15)$$

$$h_{j,t}^a = \text{Dec}(\hat{a}_{j,t-1}, s, Q, k^t, E_j), \quad (16)$$

where  $h_{j,t}^e$  and  $h_{j,t}^a$  represent the hidden state outputs of the Explanation GPT and Answer GPT, respectively.

For the iteration  $j$ , the word probability distribution of the explanation and answer can be calculated as follows:

$$p(\hat{e}_{j,t}) = \text{softmax}(W_e h_{j,t}^e + b_e), \quad (17)$$

$$p(\hat{a}_{j,t}) = \text{softmax}(W_a h_{j,t}^a + b_a). \quad (18)$$

We hypothesize that the results generated in each iteration influence model optimization. Therefore, we use the cross-entropy loss function in each iteration and then accumulate the results of each step to obtain the final loss value:

$$\mathcal{L}_{a,e}^j = - \left[ \sum_{t=1}^L \log p_{\theta}(\hat{a}_{j,t} | \hat{a}_{j,t-1}) + \sum_{t=1}^L \log p_{\theta}(\hat{e}_{j,t} | \hat{e}_{j,t-1}) + \mathcal{L}_{a,e}^{j-1} \right], \quad (19)$$

where  $\hat{a}_{j,t-1}$  and  $\hat{e}_{j,t-1}$  represent the answer and explanation words, respectively, before the decoder step  $t$  in the  $j$ -th iteration.

#### D. Cost Function

The training goal of our model is to minimize the total loss, i.e., the sum of the word-level loss of the rough answer

calculated by Equation (14) and the loss of iterative generation calculated by Equation (19). The formula for calculating this total loss is:

$$\mathcal{L}_{\text{total}} = \frac{1}{B} \sum_{i=1}^B (\lambda_1 \mathcal{L}_r + \lambda_2 \mathcal{L}_{a,e}^H), \quad (20)$$

where  $H$  and  $B$  are the total iteration number and the total number of training examples, respectively.  $\lambda_1$  and  $\lambda_2$  denote two hyperparameters.

## IV. EXPERIMENT

### A. Dataset

To demonstrate the generalization ability of our model, we conduct experiments on three datasets: the VQA-X dataset [13], the A-OKVQA dataset [14] and the VQA-X-KB dataset. Note that VQA-X-KB is the dataset we constructed by extending the VQA-X dataset, which integrates additional knowledge into the original explanations.

- 1) **VQA-X [13]**: It is a vision-language NLE dataset which contains a subset of QA pairs from the VQA v2 [33]. This approach provides explanations for the corresponding QA pairs and includes 28K images obtained from the COCO [48] dataset. We follow the official data splits of the VQA-X for our experiments, which contain 29.5k, 1.5k and 2k samples in the training set, validation set, and test set, respectively.
- 2) **A-OKVQA [14]**: In contrast to VQA-X, A-OKVQA [14] focuses on questions that require commonsense reasoning and external knowledge about the depicted scene in the images. Specifically, this database comprises 2.5K question-answer-explanation triplets, with a split of 17.1K/1.1K/6.7K for training, validation, and testing, respectively. Like in VQA-X, the images in the COCO [48] dataset were filtered to retain 23.7K images. Compared with previous VQA datasets, A-OKVQA datasets primarily include questions that require reasoning using commonsense and world knowledge with answers and explanations.
- 3) **VQA-X-KB**: This dataset is an extension of the VQA-X dataset. According to our observations, we find that explanations in the VQA-X dataset do not accurately correlate images and questions because humans often construct explanations without incorporating additional knowledge. Therefore, we extract triples from ConceptNet as additional knowledge to bridge the gap between the image and the question. The processing steps are as follows: (i) First, the faster R-CNN is utilized to extract objects from the image, and then the Stanford NER is used to extract entities from the questions and answers. (ii) Extract triples from ConceptNet with the extracted entity as the head entity, and then match the triples with the extracted objects as the tail entity to form the target triples. (iii) Similarly, we use the extracted object as the header entity to match the extracted entity to obtain the target triples. (iv) Finally, the target triples extracted from steps (ii) and (iii) are integrated into the original interpretation by splicing them and feeding the spliced

data back as new ground truth information. During this process, other contents of the dataset remain unchanged. If there is no matching target triplet, then we do not modify the original explanation in the VQA-X dataset. According to the experimental statistics, we modified 7795 (accounting for 26.4%) data points in the training set and 689 (accounting for 35.0%) data points in the test set.

### B. Experimental Details

We implement our model using PyTorch and train the model with a Tesla P100 GPU. For feature extraction, we adopt the standard off-the-shelf Oscar-Base model as the vision-language pretrained model to fuse the questions and objects in the image. ViT-B/16 [49] from the CLIP model is chosen to extract the grid feature, which denotes ViT-Base with the patch size  $16 \times 16$ . The output feature dimension  $d$  of the Oscar model and the CLIP vision encoder is 768. For knowledge retrieval, we use the Up-Down model [50] to extract 36 objects from each image and set the knowledge number  $M = 3$  in the VQA-X dataset. Like [4], we obtain the contextualized word representation from the BERT-Base model, and the output feature dimension is 768. The same parameters are used for each Distilled GPT-2 model in the iterative consensus generator. During training, we adopt the Adam optimizer [51] with an initial learning rate of  $1e-5$  to optimize the total loss function  $\mathcal{L}_{total}$ . The mini-batch size for each update is set to 36. The hyperparameters for the loss function are  $\lambda_1 = 0.3$  and  $\lambda_2 = 0.7$ .

### C. Baseline Methods and Settings

In the experiments, we compare our proposed KICNLE model and two categories of models: existing methods on the VQA-NLE task and variants of our own methods. Moreover, as mentioned before, we evaluate the performance of our method on three different datasets, i.e., VQA-X, OKVQA and VQA-X-KB.

1) *Baselines*: For the VQA-X dataset, we compare our model with six existing strong baselines.

- **PJ-X** [13] is a basic encoder-decoder framework that utilizes the multimodal compact bilinear pooling (MCB) model [31] as the vision-language encoder to predict the answer and an LSTM model as the decoder for explanation generation.
- **FME** [23] utilizes an improved Up-Down VQA model [50] with more precise image segmentation for answer inference, which avoids focusing on the background. LSTM is subsequently used for explanation generation.
- **RVT** [24] is a RATIONALE<sup>VT</sup> TRANSFORMER model that integrates multiple vision-language models to extract vision information from different granularities and feeds the visual representation, the given question and ground-truth answer into the pretrained GPT-2 [34] for explanation generation. Notably, because the question-answering part of the RVT dataset is omitted, we directly cited the

results from the e-ViL [25] benchmark, which extends RVT with BERT [52] to obtain the answer.

- **e-UG** [25] adopts the large-scale pretrained model for VQA-NLE. Specifically, it uses UNITER [19] as the encoder and GPT-2 as the decoder.
- **The NLX-GPT** [3] formulates answer prediction as a text generation task along with explanation, which uses a pretrained language model to generate answers and explanations simultaneously.
- **CaLeC** [4] utilizes chunk-level semantic information from a given question and image to infer the answer. Furthermore, it uses the lexical constraint obtained from the inference process to explicitly guide explanation generation.

For the A-OKVQA dataset, we compare our method with several existing methods [43], [44], [45], [46], [47], following the methods of Sammani et al. [3]. Specifically, Pythia [43] is an extension of [50] in which effective architectural and hyperparameter modifications are incorporated to improve the performance. ViBERT [44] and LXMERT [45] are pretrained transformer-based vision-language models. KRISP [46] incorporates implicit knowledge from a large language model (e.g., GPT-3 [53]) and explicit knowledge from a traditional knowledge graph (e.g., ConceptNet [36]) for knowledge-based VQA. Moreover, for the VQA-X-KB dataset, we focused on comparing our KICNLE model with the previous state-of-the-art NLX-GPT model [3].

2) *Variants of Our Method*: Subsequently, we compare the variants of our method on the VQA-X and A-OKVQA datasets to demonstrate the effectiveness of the different modules on the generation performance. We independently conduct corresponding ablation experiments for each module.

- **KICNLE w/o ViT**: KICNLE without the CLIP visual encoder. It only uses the region-level features extracted by the Faster R-CNN as visual input information.
- **KICNLE w/o Oscar**: KICNLE without the Oscar encoder. Only the grid-level visual features encoded by the CLIP visual encoder are utilized.
- **KICNLE w/o KRM**: KICNLE without external knowledge representations from the knowledge base. It fuses only region-level and grid-level visual features for the answer and explanation generation.
- **KICNLE  $H = n$** : KICNLE with  $n$  iterations in the iterative consensus generator, where  $n = 1, 2, 3$ .

### D. Evaluation

1) *Automatic Evaluation Metrics*: To verify the performance of each model, we utilize seven standard evaluation metrics: BLEU-(1 to 4) [54], ROUGE<sub>L</sub> [55], METEOR [56], and CIDEr [57]; these metrics are widely used in text generation tasks. All the scores are computed with the available code in [3]. In addition, we also show our VQA performance accuracy scores. Specifically, the generated answer is correct only if it is included in the set of all possible ground-truth answers.

2) *Human Evaluation Criteria*: These automatic evaluation metrics do not directly reflect the consistency between the



generated answer and explanation. Therefore, we employ five volunteers with rich educational experience to judge the quality of the same 200 samples from explanations generated by different models, as well as the consistency of the resulting answer-explanation pairs. We then average the results of the five independent evaluations [58]. The following criteria are evaluated:

- **Flu** evaluates the fluency of the generated explanation and whether it contains words or grammatical errors.
- **Rel-v** evaluates the relevance of the generated explanation and the given image.
- **Rel-q** evaluates the relevance of the generated explanation-answer pair and the given question.
- **Rel-k** reflects whether the generated explanation contains nonvisual knowledge beyond the concepts of image vision.

Specifically, **Flu** takes values from 0, 1, and 2 (higher values represent greater fluency and fewer grammatical errors), while others take a binary value.

3) *Performance Comparison*: Table I shows the automatic evaluation results of our model and the baselines on the VQA-X, A-OKVQA and VQA-X-KB datasets. Several observations are noted:

- First, our KICNLE model outperforms the previous methods on most of the seven NLG evaluation metrics in two VQA datasets: VQA-X and VQA-X-KB. Specifically, for the VQA-X dataset, KICNLE achieves the best BLEU-n (1-4), ROUGE<sub>L</sub> and CIDE<sub>r</sub> scores (e.g., +3.3% improvement in BLEU-4 and +1.8% improvement in CIDE<sub>r</sub>) while being competitive with the other SOTA methods on the METEOR in comparison. For the VQA-X-KB dataset, KICNLE also outperforms the NLX-GPT in terms of all the standard NLG evaluation metrics.
- Second, KICNLE outperforms the baselines in terms of the performance accuracy scores on the traditional VQA task. Notably, most of the baselines regard predicting the answer as a process of selecting from a given set of candidates, while NLX-GPT and KICNLE reformulate answer prediction as a text generation task. Considering this more challenging setup, KICNLE still outperforms the strongest baseline model by “0.34” points on the accuracy score in the VQA-X dataset. We further compared the accuracy scores of KICNLE and NLX-GPT, which have the same setup in the task. Specifically, KICNLE achieves an improvement in accuracy over NLX-GPT by “3.65” and “1.27” points in the VQA-X and VQA-X-KB datasets, respectively. These results support that external knowledge and the strategy of iterative generation play an important role in the VQA task.
- Third, compared with the results of the VQA-X and VQA-X-KB datasets, we observe that appending knowledge to the original explanation to form a new ground truth can improve the accuracy of the answer in both models (i.e., +2.1 and +0.59 for NLX-GPT and KICNLE, respectively). This result somewhat supports that the model generates explanations with knowledge during training and can assist the model in producing more

accurate answers. Moreover, all the NLG metrics are significantly improved, proving that models can easily generate such explanations with external knowledge.

- Finally, for the A-OKVQA dataset, our KICNLE model significantly outperforms all the baseline models in terms of answer accuracy. In particular, the accuracy of KICNLE exceeds those of KRISP and NLX-GPT by +9.97 and +3.92 respectively, in terms of accuracy. The KICNLE model retrieves more relevant knowledge from the given information (i.e., the question and image) via the KG2E [40] strategy, while KRISP simply fuses the explicit and implicit knowledge for question answering, and NLX-GPT merely utilizes the implicit knowledge that is present in the language model. Furthermore, our KICNLE model achieves the highest scores on the seven NLG metrics when compared to the NLX-GPT model. Specifically, KICNLE improves by +3.10 points on CIDE<sub>r</sub> scores, which demonstrates that external knowledge and iterative generation strategies assist in generating more faithful explanations. Moreover, these results also prove that the answer-explanation pairs generated by KICNLE are inherently consistent and can promote each other.

4) *Ablation Study*: Table I presents the results of the ablation experiment of our model on the VQA-X and A-OKVQA datasets. The results indicate that each module of our model contributes significantly to enhancing the performance of the VQA-NLE task. Based on the results of the ablation experiment, we conclude the following:

- First, the Oscar and CLIP visual encoders improve the KICNLE model, which demonstrates that both types of visual features (i.e., region-level and grid-level visual features) are beneficial for answer and explanation generation. Specifically, KICNLE w/o Oscar exhibits remarkable superiority over KICNLE w/o ViT, e.g., the accuracy of answers increases from 80.89 to 86.63 and 32.31 to 39.48 on the VQA-X and A-OKVQA datasets, respectively. This result indicates that the CLIP visual encoder can provide more robust visual features for answer and explanation generation because of the contrastive learning objective of CLIP, which results in Fusing visual and textual information is easier for downstream tasks.
- Second, KICNLE outperforms KICNLE w/o KRM on all the metrics, which demonstrates that retrieval knowledge representations obtained through KG2E are effective at improving the consistency between generated answers and explanations. Specifically, the CIDE<sub>r</sub> and accuracy results are boosted by “2.55” and “1.80” on the VQA-X dataset and by “10.46” and “1.92” on the A-OKVQA dataset.
- Third, we study the effectiveness of different numbers of iterations  $H$  in the Iterative Consensus Generator. These results demonstrate that KICNLE exhibits the best performance on VQA-X when  $H$  is 2. Additionally, KICNLE obtains competitive results on the A-OKVQA dataset when  $H$  is set to 2 or 3. These results also demonstrate that the iterative generation strategy realizes



TABLE I

MAIN AUTOMATIC METRICS RESULTS OF BASELINES AND OUR MODEL. **BOLD**: THE MAXIMUM VALUE IN THE COLUMN FOR EACH SECTION. “—” MEANS THAT THESE COMPARED MODELS MAINLY FOCUS ON THE VQA TASK RATHER THAN THE NLE TASK, SO WE CAN ONLY OBTAIN THE RESULTS OF THE ACCURACY OF THE VQA TASK

Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE <sub>L</sub>	CIDEr	Acc.
VQA-X	PJ-X [13]	57.40	42.40	30.90	22.70	19.70	46.00	82.70	76.40
	FME [23]	59.10	43.40	31.70	23.10	20.40	47.10	87.00	75.50
	RVT [24]	51.90	37.00	25.60	17.40	19.20	42.10	52.50	68.60
	e-UG [25]	57.30	42.70	31.40	23.20	22.10	45.70	74.10	80.50
	NLX-GPT [3]	64.23	49.46	37.63	28.50	<b>23.10</b>	51.51	110.60	83.07
	CaLeC [4]	61.46	47.73	36.61	28.26	21.71	50.13	100.07	86.38
	KICNLE w/o ViT	61.63	46.82	35.22	26.56	21.42	49.97	101.90	80.89
	KICNLE w/o Oscar	63.17	48.53	37.21	28.64	22.71	51.17	109.17	86.63
	KICNLE w/o KRM	63.41	48.90	37.36	28.70	22.72	51.61	110.00	84.92
	KICNLE $H = 1$	64.41	49.54	37.84	28.78	22.83	51.63	110.11	86.22
	KICNLE $H = 3$	63.46	48.92	37.35	28.54	22.82	51.42	109.80	86.69
	KICNLE $H = 2$	<b>64.91</b>	<b>49.51</b>	<b>38.07</b>	<b>29.44</b>	22.98	<b>51.71</b>	<b>112.55</b>	<b>86.72</b>
VQA-X-KB	NLX-GPT [3]	71.21	60.37	51.77	45.22	29.41	58.26	180.00	85.58
	KICNLE	<b>71.50</b>	<b>60.46</b>	<b>51.89</b>	<b>45.34</b>	<b>30.08</b>	<b>58.96</b>	<b>184.15</b>	<b>87.31</b>
A-OKVQA	Pythia [43]	-	-	-	-	-	-	-	30.60
	ViBERT [44]	-	-	-	-	-	-	-	30.60
	LXMERT [45]	-	-	-	-	-	-	-	30.70
	KRISP [46]	-	-	-	-	-	-	-	33.70
	ClipCap [47]	-	-	-	-	-	-	-	30.80
	NLX-GPT [3]	61.98	44.92	33.17	<b>24.10</b>	19.34	50.26	84.04	39.75
	KICNLE w/o ViT	59.26	41.80	30.04	21.17	18.41	47.78	75.27	32.31
	KICNLE w/o Oscar	61.25	44.10	32.20	22.16	18.61	48.93	77.07	39.48
	KICNLE w/o KRM	60.64	42.99	30.93	21.31	18.29	48.95	76.68	41.75
	KICNLE $H = 1$	60.12	43.50	31.78	22.17	18.69	49.40	80.36	43.58
	KICNLE $H = 3$	<b>63.29</b>	<b>46.04</b>	<b>33.97</b>	24.01	19.97	<b>50.75</b>	<b>87.14</b>	42.27
	KICNLE $H = 2$	62.26	45.00	33.67	<b>24.10</b>	<b>20.02</b>	50.55	84.64	<b>43.67</b>

striking performance improvement and maintains consistency between the answers and explanations. However, additional iterations do not lead to proportional improvements. This occurs because each stage relies on the results generated by the previous stages, and errors may be passed on if the answers or explanations of the previous steps are inaccurate, which reduces the accuracy of the final results. Even if the errors at each stage are small, they can increase after many iterations.

5) *Impact of  $M$  Value*: Various  $M$  values have various impacts on the results, but these impacts are not significant. The corresponding results are shown in Table II. Specifically, when  $M$  is 3, the performance is better than that with  $M$  set to 2, which demonstrates that certain effective information can be supplemented when certain knowledge is added. When  $M$  is further increased, some information redundancy will occur, resulting in worse performance with an  $M$  value of 4 than is observed with an  $M$  value of 3.

6) *Human Evaluation Results*: To ensure the reliability of our human evaluation, we first calculated the Fleiss' kappa coefficient of the evaluators for each human evaluation standard. We statistically obtained high coefficients (i.e., greater

TABLE II  
EXPERIMENTAL RESULTS OF DIFFERENT KNOWLEDGE NUMBER  $M$ . B@4, M, R AND C ARE SHORT FOR BLEU-4, METEOR, ROUGE<sub>L</sub> AND CIDEr, RESPECTIVELY

Dataset	$M$	B@4	M	R	C	Acc.
VQA-X	1	29.37	22.83	51.28	111.40	85.40
	3	<b>29.44</b>	<b>22.98</b>	<b>51.71</b>	<b>112.55</b>	<b>86.72</b>
	5	28.93	22.87	50.03	110.77	84.43
VQA-X-KB	1	<b>45.98</b>	29.60	58.19	179.97	87.10
	3	45.34	<b>30.08</b>	<b>58.96</b>	<b>184.15</b>	<b>87.31</b>
	5	43.89	27.43	54.88	172.49	85.76
A-OKVQA	1	21.49	18.62	48.51	80.45	41.05
	3	<b>24.10</b>	<b>20.02</b>	<b>50.55</b>	<b>84.64</b>	<b>43.67</b>
	5	22.02	19.76	49.71	82.38	43.48

than 0.2) for the evaluation results, which indicates that our evaluation results are reliable. Table IV lists the human evaluation results with four metrics. From these data, we conclude that:


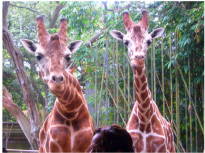

Input Image				
	(a)	(b)	(c)	(d)
Question	Is it raining?	Are these animals in a zoo?	Are the people having a party?	What event is this person celebrating most likely?
Ground Truth	Yes because the cement and blacktop are wet.	Yes because there is a barrier between the animals and people.	Yes because they are standing in a kitchen with drinks in hand and laughing.	Birthday because she is slicing a knife through a birthday cake.
NLX-GPT	Yes because the ground is wet and the sky is grey.	Yes because they are in a zoo.	No because there are no people present.	Wedding because she is cutting a cake.
CALeC	Yes because the woman is holding an umbrella.	Yes because they are in a zoo.	Yes because they are playing a game.	Birthday because she is wearing a wedding dress.
KICNLE	Yes because the ground is wet and the umbrella is open.	Yes because they are surrounded by tall fences.	Yes because there are people socializing.	Birthday because she is cutting a birthday cake.
ConceptNet	[[Umbrella]] is used for [[rainy days]]	[[Fence]] is used for [[containing animals]]	[[Person]] is used for [[social interaction]]	[[Cake]] is related to [[birthday]]

Fig. 3. Case study of sample output answers and explanations generated by NLX-GPT, CALeC, and our KICNLE model.

TABLE III

EXPERIMENTAL RESULTS OF OUR KICNLE AND CHATGPT. B@4 AND M ARE SHORT FOR BLEU-4 AND METEOR, RESPECTIVELY

Dataset	Model	B@4	M	Rel-q	Acc.
VQA-X	ChatGPT	19.80	20.54	<b>0.79</b>	70.05
	KICNLE	<b>26.82</b>	<b>22.40</b>	0.75	<b>84.00</b>
VQA-X-KB	ChatGPT	33.58	29.27	0.79	76.50
	KICNLE	<b>40.81</b>	<b>32.80</b>	<b>0.82</b>	<b>86.50</b>
A-OKVQA	ChatGPT	16.08	14.64	<b>0.69</b>	<b>39.50</b>
	KICNLE	<b>22.83</b>	<b>19.01</b>	0.66	38.00

TABLE IV

HUMAN EVALUATION RESULTS OF BASELINES AND OUR MODELS. **BOLD**: THE MAXIMUM VALUE IN THE COLUMN

Method	Flu	Rel-v	Rel-q	Rel-k
e-UG	2.82	0.57	0.41	0.32
NLX-GPT	3.26	0.54	0.50	0.39
CaLeC	3.63	0.65	0.63	0.25
KICNLE	<b>4.09</b>	<b>0.79</b>	<b>0.73</b>	<b>0.64</b>

- For **Flu**, the evaluation result is e-UG < NLX-GPT < CaLeC < KICNLE, which shows that our KICNLE is adept at generating more fluent and smooth sentences.
- The indicator **Rel-v** reflects that the correlation between the generated answer-explanation pair of our model and the given image is greater than that of the baseline models. This is attributed to the utilization of two types of visual features (i.e., region-level and grid-level features),

which enable our model to capture more accurate visual scenes about the image for answer and explanation generation.

- The **Rel-q** score indicates that the relevance of the given question and the answer-explanation pair generated by our model surpasses that of the baseline methods. Since we utilize Oscar for question and region-level visual feature alignment and design a question-aware encoder in the KRM module for knowledge selection, the KICNLE model is able to capture more relevant visual and knowledge representations for the VQA-NLE task.
- The results generated by our model are better than those of the baselines for the **Rel-k** score because only our KICNLE model explicitly makes use of external knowledge for the answer and explanation generation.

### E. Case Study

As shown in Figure 3, we show some cases in which the results were generated from the NLX-GPT, CaLeC and our KICNLE models on the VQA-X dataset. We find that:

- **Knowledge.** Compared with the NLX-GPT and CaLeC models, our model is able to integrate more accurate knowledge bridging the gap between images and questions for explanation generation, such as “Fence is used for containing animals” in Patient (b). The introduction of additional knowledge can make the generated explanations more reasonable.
- **Relevance.** Case (c) exemplifies this, where our model generates the answer “Yes” along with the explanation “Someone is socializing”, which is related to the image content of “the people having a party”. The generated answers also tend to be more accurate. However, the compared NLX-GPT model generates a “no people


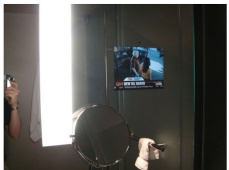
Input Image		
	(a)	(b)
Question	What kind of plant is the big zebra eating?	What room is this?
Ground Truth	Grass because it is long and brown and is what grass looks like in the African plains.	Bathroom because there is a towel hanging on the shower door.
NLX-GPT	Grass because it is tall and has a long tail.	Bathroom because there is a sink.
CALeC	Grass because it is green and has a long neck.	Bathroom because there is a toilet and a sink.
KICNLE	Grass because it is tall and green.	Bathroom because there is a mirror and a sink.
ConceptNet	[[Grass]] is used for [[containing animals]]	[[Mirror]] can be found in the [[bathroom]]

Fig. 4. Bad case study of sample output answers and explanations generated by NLX-GPT, CALeC, and our KICNLE model.

present” explanation, which is irrelevant to the image content.

- **Consistency.** After multiple iterations, our model achieves improved consistency between the generated answer and explanation. As shown in Case (d), when the generated answer is “birthday”, our model can generate birthday-related explanations, i.e., cutting the birthday cake. The answers and explanations generated by the compared models (i.e., NLX-GPT and CALeC) are not consistent and are related to birthdays and weddings, respectively.

Furthermore, Figure 4 shows several bad cases generated from our KICNLE, NLX-GPT and CaLeC models. In these cases, the generated explanations are different from the ground truth. We find that the additional knowledge in cases (a) and (b) helps our model generate correct answers, i.e., establish a connection between the grass and the animal zebra in Case (a) and between the mirror and the bathroom in Case (b). However, the generated explanation process is affected by some cooccurrences (e.g., most grass is green, or there is a sink in the toilet in the dataset), which results in incorrect generation of the three models. Thus, we can try to eliminate the bias caused by the dataset and generate more accurate explanations in future work.

## V. IMPACT OF GPT VERSIONS

The GPT has different versions, and the performance disparities among the various versions of GPTs are substantial. Thus, in this section, we discuss the impact of GPT2 and GPT3.5 (i.e., ChatGPT) on framework performance. GPT3.5 adds text on image captioning based on the prompt design of GPT2, which enables GPT3.5 to also capture image information. The captioning text of the image is derived from human annotation.

For each sample, we generate results by calling ChatGPT’s API each time. We have included a complete and correct example in the prompt design, including image description, questions, answers, and explanations, which helps ChatGPT learn relevant generation patterns. Specifically, we test on 200 randomly selected samples, and the results are shown in Table III.

We find that GPT3.5 failed to significantly improve the automatic evaluation indicators of our framework, mainly because (i) the caption may not reflect the complete image information, even if it is manually annotated; and (ii) GPT3.5 does not perform training, which prevents it from generating text content similar to ground truth.

In addition, GPT3.5 performs better than GPT2 on the manual evaluation indicator rel-q. This finding shows that GPT3.5 performs well in VQA and NLE tasks and can generate answers and explanations that are different from ground truths but reasonable. This also shows that our framework has great potential to produce better results when replacing the more powerful GPT.

## VI. CONCLUSION

In this paper, we highlight two distinct challenges associated with VQA-NLE tasks, i.e., consistency and correlation. To address the first challenge, we propose a model that takes each other into account in generating answers and explanations through a multi-iterative generation method to generate consistent answers and explanations. For the second challenge, we seamlessly integrate knowledge into a pretrained language model to bridge the gaps between images and questions so that the generated explanations contain accurate relations. To obtain the target knowledge, we first extract candidate knowledge related to visual objects from ConceptNet and then



directly retrieve the knowledge that is most relevant to the given question. Furthermore, we constructed a new dataset called VQA-X-KB based on the VQA-X dataset. Specifically, this dataset incorporates the most relevant knowledge related to the explanation, which is appended after the original explanation as the new ground truth. The experimental results show that our model is superior to the existing methods in terms of automatic and human evaluation metrics on the VQA-X, VQA-X-KB, and A-OKVQA datasets, establishing its effectiveness in generating high-quality explanations and answers.

## REFERENCES

- [1] A. A. Yusuf, F. Chong, and M. Xianling, "An analysis of graph convolutional networks and recent datasets for visual question answering," *Artif. Intell. Rev.*, vol. 55, no. 8, pp. 6277–6300, Dec. 2022.
- [2] W. Guo, Y. Zhang, J. Yang, and X. Yuan, "Re-attention for visual question answering," *IEEE Trans. Image Process.*, vol. 30, pp. 6730–6743, 2021.
- [3] F. Sammani, T. Mukherjee, and N. Deligiannis, "NLX-GPT: A model for natural language explanations in vision and vision-language tasks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 8312–8322.
- [4] Q. Yang, Y. Li, B. Hu, L. Ma, Y. Ding, and M. Zhang, "Chunk-aware alignment and lexical constraint for visual entailment with natural language explanations," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 3587–3597.
- [5] K. Kafle and C. Kanan, "Visual question answering: Datasets, algorithms, and future challenges," *Comput. Vis. Image Understand.*, vol. 163, pp. 3–20, Oct. 2017.
- [6] J. Yuan et al., "Discriminative style learning for cross-domain image captioning," *IEEE Trans. Image Process.*, vol. 31, pp. 1723–1736, 2022.
- [7] H. Liu, S. Zhang, K. Lin, J. Wen, J. Li, and X. Hu, "Vocabulary-wide credit assignment for training image captioning models," *IEEE Trans. Image Process.*, vol. 30, pp. 2450–2460, 2021.
- [8] Y. Huang, J. Chen, W. Ouyang, W. Wan, and Y. Xue, "Image captioning with end-to-end attribute detection and subsequent attributes prediction," *IEEE Trans. Image Process.*, vol. 29, pp. 4013–4026, 2020.
- [9] M. Yang et al., "An ensemble of generation- and retrieval-based image captioning with dual generator generative adversarial network," *IEEE Trans. Image Process.*, vol. 29, pp. 9627–9640, 2020.
- [10] A. C. Butler, N. Godbole, and E. J. Marsh, "Explanation feedback is better than correct answer feedback for promoting transfer of learning," *J. Educ. Psychol.*, vol. 105, no. 2, pp. 290–298, May 2013.
- [11] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. ICML*, in Proceedings of Machine Learning Research, vol. 139, M. Meila and T. Zhang, Eds., 2021, pp. 8748–8763.
- [12] X. Li et al., "OSCAR: Object-semantics aligned pre-training for vision-language tasks," in *Proc. ECCV*, in Lecture Notes in Computer Science, vol. 12375, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds. Cham, Switzerland: Springer, 2020, pp. 121–137.
- [13] D. H. Park et al., "Multimodal explanations: Justifying decisions and pointing to the evidence," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8779–8788.
- [14] D. Schwenk, A. Khandelwal, C. Clark, K. Marino, and R. Mottaghi, "A-OKVQA: A benchmark for visual question answering using world knowledge," in *Proc. ECCV*, in Lecture Notes in Computer Science, vol. 13668. Cham, Switzerland: Springer, 2022, pp. 146–162.
- [15] H. Xu and K. Saenko, "Ask, attend and answer: Exploring question-guided spatial attention for visual question answering," in *Proc. ECCV*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 451–466.
- [16] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, "Deep modular co-attention networks for visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6281–6290.
- [17] J. Wu and R. J. Mooney, "Entity-focused dense passage retrieval for outside-knowledge visual question answering," 2022, *arXiv:2210.10176*.
- [18] Y. Guo, L. Nie, Y. Wong, Y. Liu, Z. Cheng, and M. Kankanhalli, "A unified end-to-end retriever-reader framework for knowledge-based VQA," in *Proc. 30th ACM Int. Conf. Multimedia*, J. Magalhães, et al., Eds., Oct. 2022, pp. 2061–2069.
- [19] Y. Chen et al., "UNITER: Universal image-text representation learning," in *Proc. ECCV*, in Lecture Notes in Computer Science, vol. 12375, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds. Cham, Switzerland: Springer, 2020, pp. 104–120.
- [20] Z. Huang, Z. Zeng, Y. Huang, B. Liu, D. Fu, and J. Fu, "Seeing out of the box: End-to-end pre-training for vision-language representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12976–12985.
- [21] J. Li, D. Li, C. Xiong, and S. C. H. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. ICML*, in Proceedings of Machine Learning Research, vol. 162, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, Eds., 2022, pp. 12888–12900.
- [22] H. A. Ayyubi, M. Mehrab Tanjim, J. J. McAuley, and G. W. Cottrell, "Generating rationales in visual question answering," 2020, *arXiv:2004.02032*.
- [23] J. Wu and R. J. Mooney, "Faithful multimodal explanation for visual question answering," in *Proc. ACL Workshop*, T. Linzen, G. Chrupala, Y. Belinkov, and D. Hupkes, Eds., 2019, pp. 103–112.
- [24] A. Marasovic, C. Bhagavatula, J. S. Park, R. L. Bras, N. A. Smith, and Y. Choi, "Natural language rationales with full-stack visual reasoning: From pixels to semantic frames to commonsense graphs," in *Proc. EMNLP Finding*, T. Cohn, Y. He, and Y. Liu, Eds., 2020, pp. 2810–2829.
- [25] M. Kayser et al., "E-ViL: A dataset and benchmark for natural language explanations in vision-language tasks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 1224–1234.
- [26] X. Chen and A. Gupta, "Spatial memory for context reasoning in object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4106–4116.
- [27] P. Barbiero et al., "Interpretable neural-symbolic concept reasoning," in *Proc. ICML*, in Proceedings of Machine Learning Research, vol. 202, 2023, pp. 1801–1825.
- [28] J. Mao, C. Gan, P. Kohli, J. B. Tenenbaum, and J. Wu, "The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision," in *Proc. ICLR*, 2019, pp. 1–13.
- [29] K. Yi, J. Wu, C. Gan, A. Torralba, P. Kohli, and J. Tenenbaum, "Neural-symbolic VQA: Disentangling reasoning from vision and language understanding," in *Proc. NeurIPS*, S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., 2018, pp. 1039–1050.
- [30] C. Gan, Y. Li, H. Li, C. Sun, and B. Gong, "VQS: Linking segmentations to questions and answers for supervised attention in VQA and question-focused semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1829–1838.
- [31] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," in *Proc. EMNLP*, J. Su, X. Carreras, and K. Duh, Eds., 2016, pp. 457–468.
- [32] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
- [33] S. Antol et al., "VQA: Visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2425–2433.
- [34] A. Radford et al., "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.
- [35] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. NeurIPS*, 2015, pp. 91–99.
- [36] R. Speer, J. Chin, and C. Havasi, "ConceptNet 5.5: An open multilingual graph of general knowledge," in *Proc. 31st AAAI Conf. Artif. Intell.*, S. Singh and S. Markovitch, Eds., 2017, pp. 4444–4451.
- [37] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Found. Trends Inf. Retr.*, vol. 3, no. 4, pp. 333–389, 2009.
- [38] J. Martineau and T. Finin, "Delta TFIDF: An improved feature space for sentiment analysis," in *Proc. International AAAI Conf. Web Social Media*, vol. 3, no. 1, 2009, pp. 258–261.
- [39] J. Xie, W. Fang, Y. Cai, Q. Huang, and Q. Li, "Knowledge-based visual question generation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 11, pp. 7547–7558, Nov. 2022.

- [40] S. He, K. Liu, G. Ji, and J. Zhao, "Learning to represent knowledge graphs with Gaussian embedding," in *Proc. 24th ACM Int. Conf. Inf. Knowl. Manag.*, 2015, pp. 623–632.
- [41] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," 2019, *arXiv:1910.01108*.
- [42] J. Xie, N. Peng, Y. Cai, T. Wang, and Q. Huang, "Diverse distractor generation for constructing high-quality multiple choice questions," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 30, pp. 280–291, 2022.
- [43] Y. Jiang, V. Natarajan, X. Chen, M. Rohrbach, D. Batra, and D. Parikh, "Pythia v0.1: The winning entry to the VQA challenge 2018," 2018, *arXiv:1807.09956*.
- [44] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019, pp. 13–23.
- [45] H. Tan and M. Bansal, "LXMERT: Learning cross-modality encoder representations from transformers," in *Proc. EMNLP*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds., 2019, pp. 5099–5110.
- [46] K. Marino, X. Chen, D. Parikh, A. Gupta, and M. Rohrbach, "KRISP: Integrating implicit and symbolic knowledge for open-domain knowledge-based VQA," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14106–14116.
- [47] R. Mokady, A. Hertz, and A. H. Bermano, "ClipCap: CLIP prefix for image captioning," 2021, *arXiv:2111.09734*.
- [48] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. ECCV*. Springer, 2014, pp. 740–755.
- [49] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021, pp. 1–12.
- [50] P. Anderson et al., "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6077–6086.
- [51] D. P. Kingma and J. Ba, "ADAM: A method for stochastic optimization," in *Proc. ICLR*, 2015, pp. 1–11.
- [52] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL*, J. Burstein, C. Doran, and T. Solorio, Eds., 2019, pp. 4171–4186.
- [53] T. B. Brown et al., "Language models are few-shot learners," in *Proc. NeurIPS*, vol. 33, 2020, pp. 1877–1901.
- [54] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. ACL*, 2002, pp. 311–318.
- [55] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Proc. ACL Workshop*, 2004, pp. 74–81.
- [56] M. J. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *Proc. ACL Workshop*, 2014, pp. 376–380.
- [57] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4566–4575.
- [58] C. Xing et al., "Topic aware neural response generation," in *Proc. AAAI*, 2017, pp. 3351–3357.



**Jiayuan Xie** received the Ph.D. degree from South China University of Technology in 2023. He is currently a Postdoctoral Research Fellow with the School of Computing, The Hong Kong Polytechnic University. His research interests include multimedia and question generation.



**Yi Cai** (Member, IEEE) received the Ph.D. degree in computer science from The Chinese University of Hong Kong. He is currently a Professor with South China University of Technology (SCUT). His research works are published in many conferences and journals, such as IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, *Neural Networks*, *Knowledge-Based Systems*, *Engineering Applications of Artificial Intelligence*, and *Neurocomputing*, and AAAI, COLING, CIKM, AAMAS, DASFAA, and other international conferences about perspective mining, cognitive modeling, information retrieval, and semantic analysis. He received the National Science and Technology Academic Publications Fund. His two books are published by the Higher Education Press and Springer Press Monograph. His research interests include recommendation systems, personalized search, semantic web, and data mining. At the same time, he has served as a reviewer for several important international academic journals and international academic conferences, such as IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, *ACM Transactions on Internet Technology*, IEEE INTELLIGENT SYSTEMS, *Information Science*, *Knowledge-Based Systems*, IJCAI, AAAI, COLING, CIKM, and DASFAA.



**Jiali Chen** received the bachelor's degree from South China University of Technology, Guangdong, China, in 2023, where he is currently pursuing the Ph.D. degree. His research interests include multimedia, sentiment analysis, and question generation.



**Ruohang Xu** is currently pursuing the B.E. degree with the School of Software Engineering, South China University of Technology. His research interests include multimedia.



**Jiexin Wang** (Member, IEEE) received the B.Eng. degree in computer science from South China University of Technology (SCUT), Guangzhou, and the M.Sc. and Ph.D. degrees in informatics from Kyoto University, Japan. He is currently an Associate Research Professor with SCUT. His research interests include information retrieval and natural language processing, including question answering, question generation, time series analysis, and pre-trained language models.



**Qing Li** (Fellow, IEEE) received the B.Eng. degree in computer science from Hunan University, Changsha, and the M.Sc. and Ph.D. degrees in computer science from the University of Southern California, Los Angeles. He is currently a Chair Professor with the Department of Computing, The Hong Kong Polytechnic University. His research interests include multi-modal data management, conceptual data modeling, social media, web services, and e-learning systems. He has authored or coauthored more than 400 publications in these areas. He is actively involved in the research community. He is a fellow of IEEE/IET, U.K., and a Distinguished Member of CCF, China. He served as the conference chair and the program chair/co-chair for numerous major international conferences. He also sits on the Steering Committees of DASFAA, ER, ACM RecSys, IEEE U-MEDIA, and ICWL. He has served as an Associate Editor for several major technical journals, including IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, *ACM Transactions on Internet Technology*, *Data Science and Engineering*, *World Wide Web*, and the *Journal of Web Engineering*.