

# Generalizing Visual Question Answering from Synthetic to Human-Written Questions via a Chain of QA with a Large Language Model

Taehee Kim<sup>a</sup>, Yeongjae Cho<sup>b</sup>, Heejun Shin<sup>a</sup>, Yohan Jo<sup>b</sup> and Dongmyung Shin<sup>a,\*</sup>

<sup>a</sup>Radisen Co. Ltd.

<sup>b</sup>Seoul National University

**Abstract.** Visual question answering (VQA) is a task where an image is given, and a series of questions are asked about the image. To build an efficient VQA algorithm, a large amount of QA data is required which is very expensive. Generating synthetic QA pairs based on templates is a practical way to obtain data. However, VQA models trained on those data do not perform well on complex, human-written questions. To address this issue, we propose a new method called *chain of QA for human-written questions* (CoQAH). CoQAH utilizes a sequence of QA interactions between a large language model and a VQA model trained on synthetic data to reason and derive logical answers for human-written questions. We tested the effectiveness of CoQAH on two types of human-written VQA datasets for 3D-rendered and chest X-ray images and found that it achieved state-of-the-art accuracy in both types of data. Notably, CoQAH outperformed general vision-language models, VQA models, and medical foundation models with no finetuning. The source code for CoQAH is available [10].

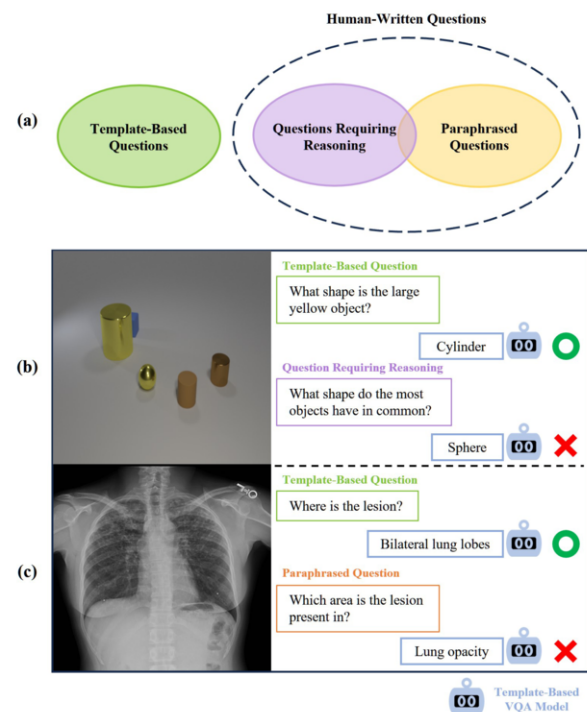
## 1 Introduction

Visual question answering (VQA) aims to build an automated algorithm to answer a series of questions regarding a given image. This task has a wide range of potential applications, such as interpreting medical images [14] and supporting visually impaired people [3].

Recently, general vision language models (VLMs), trained using a massive amount of image and text data, have shown promising results in solving VQA tasks [2, 1, 16]. However, their performance is limited when tested in VQA tasks of specific domains (e.g., VLM model trained on large amounts of medical images and captions on the web vs. VQA model specialized for chest X-rays) [13].

This challenge highlights the need to train or finetune VLM models using VQA data in a particular domain. However, acquiring such data is challenging, requiring experts to design VQA datasets carefully without contradiction and logical errors. For instance, to create a VQA dataset of chest X-rays, radiologists must write QA pairs consistent with the given radiographs [12, 15].

To address this challenge, a template-based approach, which automatically synthesizes QA pairs based on pre-defined templates, has been adopted. This approach produced high-quality VQA datasets with minimum errors (e.g., CLEVR [7] and MIMIC-Diff-VQA [5]). Based on these datasets, some studies reported very high accuracy on



**Figure 1.** (a) Human-written questions, compared to the fixed format of the template-based questions, include more complex free-form questions, such as ones requiring reasoning. (b), (c) Example cases where a template-based VQA model fails to answer human-written questions correctly.

internal testing (e.g., 99.7% in CLEVR) [5, 6, 19, 9]. However, they also found that those VQA models failed to answer human-written questions that deviated from the templates, such as questions requiring complex reasoning (e.g., *what shape is the large yellow object?* vs. *what shape do the most objects have in common?* in Figure 1b) and, even paraphrased questions (e.g., *where is the lesion?* vs. *which area is the lesion present in?* in Figure 1c).

In this study, we propose a novel method called *chain of QA for human-written questions* (CoQAH) which can correctly answer complex, human-written questions beyond a fixed set of synthetic, template-based questions without the need for finetuning. In the core

\* Corresponding Author. Email: shinsae11@radisentech.com.

part of CoQAH, a large language model (LLM) sequentially asks template-based questions to a VQA model trained with synthetic data. This step-by-step QA process enables the LLM to collect valuable and accurate information about an image and reach a reasonable conclusion for a given question.

To check the effectiveness of CoQAH, we utilized two different types of template-based VQA datasets (CLEVR for 3D-rendered images and MIMIC-Diff-VQA for chest X-rays) for AI training. We then used multiple human-written VQA datasets (CLEVR-Human [8], VQA-RAD [12], and SLAKE [15]) for AI testing. As a result, CoQAH achieved state-of-the-art performance on all the human-written VQA datasets, surpassing general VLMs and other VQA models.

Our contributions are summarized as follows: (1) we propose CoQAH, a new method that answers human-written questions through a step-by-step QA process between an LLM and a VQA model. (2) our method achieved state-of-the-art performance on multiple human-written VQA datasets, surpassing general VLMs and other VQA models. (3) we demonstrated the effectiveness of CoQAH on two types of VQA datasets, including 3D-rendered images (CLEVR and CLEVR-Human) and chest X-rays (MIMIC-Diff-VQA, VQA-RAD, and SLAKE).

## 2 Related Works

### 2.1 General Vision Language Models

A general VLM is an AI model that can understand images and texts, along with their relationships, in various tasks. One of the earliest approaches to creating a general VLM was to combine separate vision and text encoders, aligning features from both encoders [20, 26]. Another approach utilized a pre-trained LLM, which can understand the meaning and context of texts. Tsimpoukelli et al. [22] proposed training and combining a vision encoder with a frozen LLM. Similarly, Alayrac et al. [2] proposed finetuning the connecting layers between an LLM and a vision encoder. More recently, Liu et al. [16] proposed training VLMs using the instructions of various tasks to make a general-purpose assistant. These general VLMs can be applied to various vision-language tasks, but they often lack expertise in specific domains. CoQAH addresses this issue by employing a specialized VQA model (e.g., chest X-ray VQA) that interacts with an LLM.

### 2.2 Template-Based VQA

Creating VQA datasets manually is a time-consuming and expensive process. To overcome this challenge, some researchers have attempted to synthesize QA pairs using pre-defined templates [5, 7]. For instance, one study introduced a VQA dataset for chest X-rays by extracting abnormal findings from radiologist's reports [5]. Similarly, another study created 3D-rendered images of various objects and generated complex template-based questions related to the visual features of those images [7]. Several other studies have used such datasets to train and evaluate the visual understanding capabilities of AI models [6, 19, 9, 5].

### 2.3 LLM-Aided VQA

Several studies have used an LLM to solve VQA problems [24, 21, 4, 29, 25]. This is done by feeding the relevant information to the LLM, which then uses its high contextual understanding and reasoning ability to answer the question. For example, some researchers

[24, 21, 4] have proposed feeding an image description from a captioning model to an LLM, while others have utilized VQA and captioning models together to give visual clues to an LLM [29]. Recently, some researchers have suggested using iterative interactions between an LLM and a VQA model to answer complex visual questions [25]. CoQAH uses a template-based VQA model that is specialized to questions in specific domains, which makes it different from other methods above.

## 3 Method

### 3.1 CoQAH

The CoQAH method aims to conclude a reasonable answer for a given complex, human-written question (i.e., user question; see Figure 2a) by employing synthetic, template-based VQA data only. To achieve this, it utilizes interactions between two sub-models, a template-based VQA model ( $VQA_t$ ; see Supplementary Table 1 and 2 for the performance of VQA models [11]) and a large language model ( $LLM$ ).

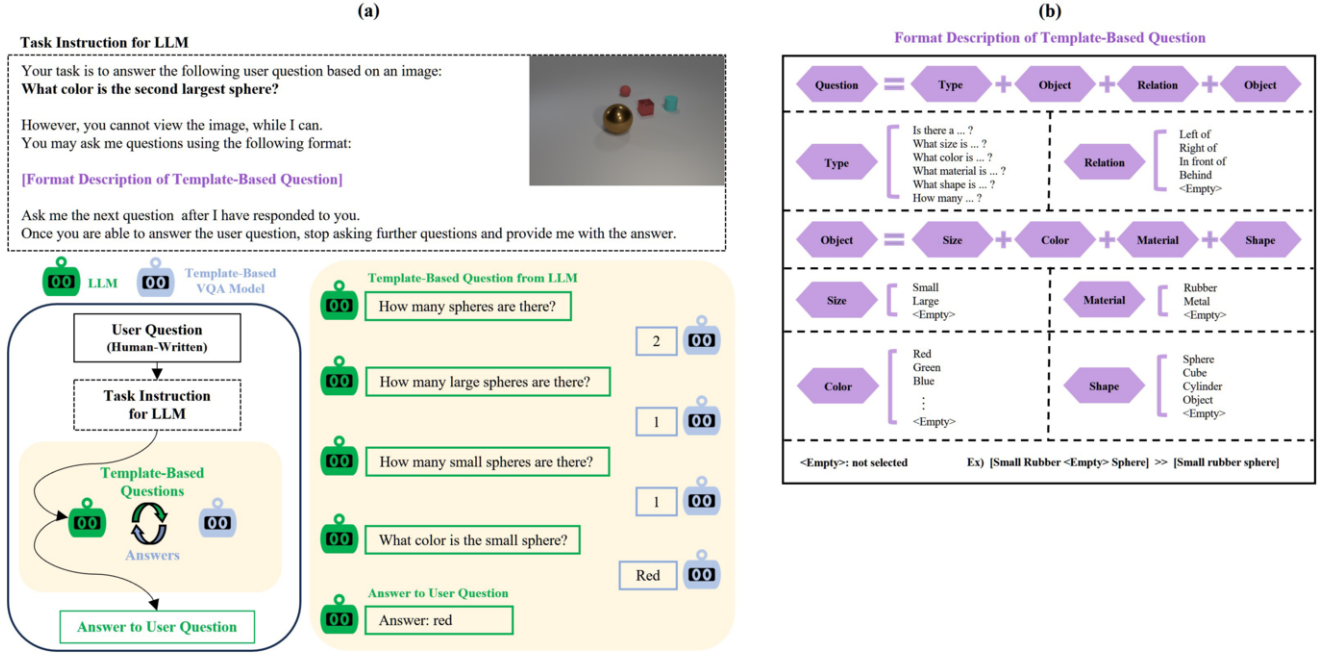
Suppose a user question and an image from a human-written VQA dataset ( $q_h \in Q_h, i_h \in I_h; Q_h$  for human-written questions and  $I_h$  for images) are given; first, the  $LLM$  is prompted with a task instruction ( $\rho$ ; see Section 3.2 for detail) that directs the model to generate a template-based question that conforms to the question templates used to train the VQA model ( $q_t^1 = LLM(\rho)$ ;  $q_t^1 \in Q_t$  where  $Q_t$  includes all possible template-based questions; see Section 3.2). Next,  $VQA_t$  answers the question ( $a_t^1 = VQA_t(q_t^1, i_h)$ ). After that, all the previous dialogue, including the task instruction, template-based questions, and answers from  $VQA_t$  ( $\rho + q_t^1 + a_t^1$ ), is again fed into  $LLM$ , generating a next template-based question ( $q_t^2 = LLM(\rho, q_t^1, a_t^1)$ ). This QA process is repeated until  $LLM$  has sufficient information to reach the final answer for the user question or it reaches the maximum number of questions to be queried (twenty questions for CLEVR-Human and five questions for VQA-RAD and SLAKE; see Section 5.4). Note that  $LLM$  cannot access the image, whereas  $VQA_t$  does.

### 3.2 Task Instruction

The task instruction  $\rho$  is a detailed description of how  $LLM$  should respond to and interact with  $VQA_t$  (Figure 2a). The instruction starts with *Your task is to answer the following user question based on an image:* followed by the user question. Then, we included a statement to declare that  $LLM$  cannot see a given image (*However, you cannot view the image, while I can.*). After that, we asked  $LLM$  to give a question that can only be generated based on a template format (*You may ask me questions using the following format;* see Section 3.3 for details). In the last part, we added some sentences to ask  $LLM$  to query the next question step-by-step after having the response from  $VQA_t$  (*Ask me the next question after I have responded to you.*) and conclude the answer for the user question whenever it is ready (*Once you are able to answer the user question, stop asking further questions and provide me with the answer.*).

### 3.3 Template-Based Questions from LLM

Figure 2b describes how we instructed  $LLM$  to generate a template-based question for the CLEVR dataset [7]. We first let  $LLM$  know the structure of the question, which is composed of several different entities ( $\langle question \rangle = \langle type \rangle + \langle object \rangle + \langle relation \rangle + \langle object \rangle$  in Figure 2b). Subsequently, we asked  $LLM$  to choose one of the



**Figure 2.** An overview of the proposed CoQAH method. (a) An example of a task instruction is shown at the top. The figure on the left describes an overall interaction process between an LLM and a template-based VQA model to reach the final answer to the user question. On the right, the figure represents an example of dialogue between the two models. (b) The template format for the questions in the CLEVR dataset is described, which is composed of several different entities ( $\langle question \rangle = \langle type \rangle + \langle object \rangle + \langle relation \rangle + \langle object \rangle$ ). For each entity, a few options are available to be selected (e.g., small or large or  $\langle Empty \rangle$  for  $\langle Size \rangle$  entity).

options available for each entity (e.g., small or large or  $\langle Empty \rangle$  for  $\langle Size \rangle$  entity). The template format for MIMIC-Diff-VQA is shown in Supplementary Figure 1. The detailed description of the task instructions for both CLEVR and MIMIC-Diff-VQA is displayed in Supplementary Figure 2 and 3.

### 3.4 Existence and Uniqueness Handler

Even if *LLM* successfully generated a template-based question, in some cases, we observed that there were still some questions that had logical errors (see Figure 3): questions that ask the property of an object that does not exist in the image (existence violation in Figure 3b) and ask the property of multiple objects with the same condition (uniqueness violation in Figure 3c). For these types of questions,  $VQA_t$  provided incorrect answers unexpectedly.

To address this problem, we introduced an existence and uniqueness handler (EUH), which continually checks the violation of both conditions on a template-based question generated from *LLM*. As described in Algorithm 1, we first extracted the object entity ( $\langle object \rangle + \langle relation \rangle + \langle object \rangle$ ) from the question ( $\langle type \rangle + \langle object \rangle + \langle relation \rangle + \langle object \rangle$ ). For example, the object entity ("small object" + "right of" + "shiny red object") was extracted from the question (What color is the small object right of the shiny red object?). Then, we checked the existence of the object entity by querying to  $VQA_t$  (e.g., Is there a small object right of the shiny red object?). When the existence of the object was confirmed, we continued to check the uniqueness by asking  $VQA_t$  to count the number of objects (e.g., How many small objects right of the shiny red object are there?). Using EUH, we prevented *LLM* from directly providing a

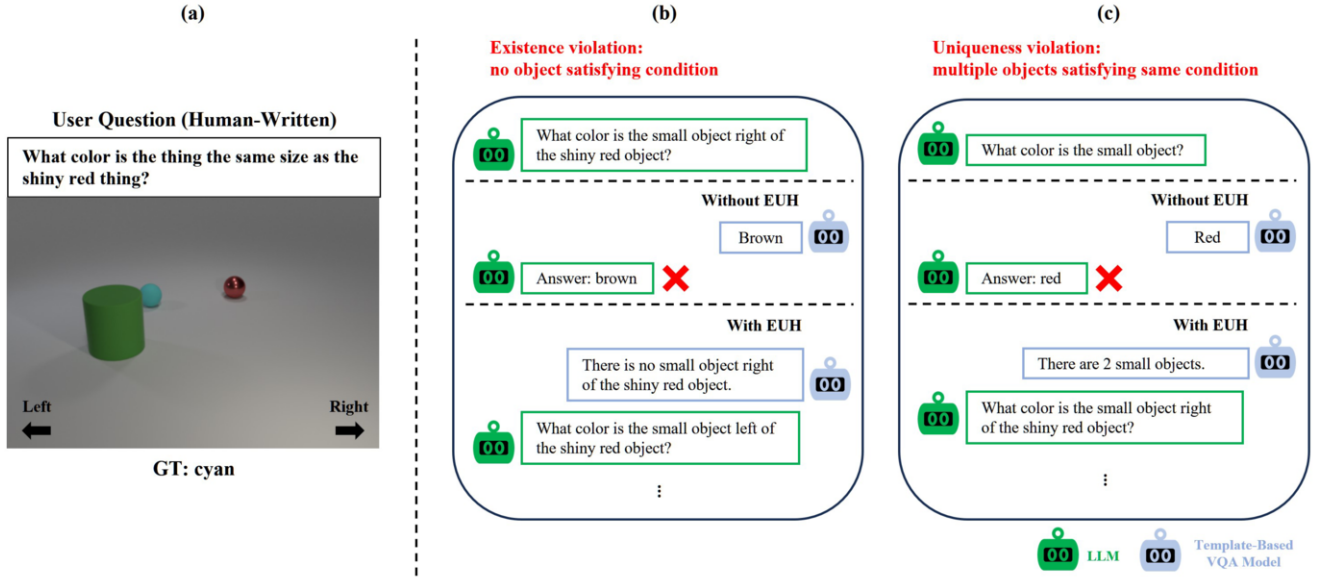
wrong answer without considering logical contradiction.

#### Algorithm 1 Existence and uniqueness handler

**Input:** Question from LLM  $Q_L$ ; Input image  $I$ ; VQA model  $VQA(\cdot)$ ; A function that extracts  $\langle Object \rangle + \langle Relation \rangle + \langle Object \rangle$  part from the question  $\langle Type \rangle + \langle Object \rangle + \langle Relation \rangle + \langle Object \rangle$   $Extract\_Entity(\cdot)$

**Output:** Answer to  $Q_L$   $A_L$

- 1:  $E \leftarrow Extract\_Entity(Q_L)$
- 2:  $Q_E \leftarrow \text{"Is there a"} + E + \text{"?"}$
- 3:  $A_E \leftarrow VQA(Q_E, I)$
- 4: **if**  $A_E$  is "No" **then**
- 5:    $A_L \leftarrow \text{"There is no"} + E$
- 6: **else if**  $A_E$  is "Yes" **then**
- 7:    $Q_U \leftarrow \text{"How many"} + E + \text{"are there?"}$
- 8:    $A_U \leftarrow VQA(Q_U, I)$
- 9:   **if**  $A_U$  is not "1" **then**
- 10:      $A_L \leftarrow \text{"There are"} + A_U + E$
- 11:   **else if**  $A_U$  is "1" **then**
- 12:      $A_L \leftarrow VQA(Q_L, I)$
- 13:   **end if**
- 14: **end if**
- 15: **Return**  $A_L$



**Figure 3.** Illustration of how the existence and uniqueness handler (EUH) can prevent an LLM from concluding an incorrect answer for a given user question. (a) An example case includes a question, an answer, and an image. (b) A dialogue between an LLM and a VQA model when the existence of an object is violated. With EUH, the VQA model checks the presence of an object and lets LLM know its existence. (c) A dialogue when the uniqueness of an object is violated. With EUH, the VQA model successfully lets LLM know the number of objects satisfying the same condition.

## 4 Experiment Setup

### 4.1 Datasets

For datasets, we collected two types of images (and question-answer pairs), including 3D-rendered images and chest X-ray images, to show the effectiveness of CoQAH in different image types. We first trained VQA models using template-based datasets (CLEVR [7] for 3D-rendered and MIMIC-Diff-VQA [5] for chest X-ray) and externally validated the model’s performance on the human-written datasets (CLEVR-Human [8] for 3D-rendered; VQA-RAD [12] and SLAKE [15] for chest X-ray).

**CLEVR** is a synthetic VQA dataset (70,000 image and QA pairs for training, 15,000 for validation, and 15,000 for testing). 3D objects of different sizes, colors, materials, and shapes were rendered for each image. Then, questions were generated based on a template-based algorithm, and the answers for those questions were automatically selected among a set of candidates (e.g., 3, blue, cylinder, etc.).

**CLEVR-Human** is a variant of the CLEVR dataset that replaced questions and answers in some images of CLEVR using human questioners (17,817 for training; 7,202 for validation). The questioners created new human-written questions without any format restriction, while the answers for the questions were chosen among the same candidates in CLEVR. Currently, a test set is not available. Therefore, we reported AI performance using the validation set.

**MIMIC-Diff-VQA** is a chest X-ray VQA dataset. QA pairs were generated using an automatic template algorithm based on radiologists’ reports (700,703 QA pairs from 164,324 reports and images; 80% for training, 10% for validation, and 10% for testing).

**VQA-RAD and SLAKE** are medical image VQA datasets where clinical trainees (for VQA-RAD) or physicians (for SLAKE) wrote human-written questions and answers manually for each image. We

used only the chest X-rays in those datasets (107 images for VQA-RAD and 179 images for SLAKE) for AI evaluation. In both datasets, answers are categorized as either closed-form or open-form: the answers of the closed-form questions were required to be one between two or three options (e.g., yes or no; 511 questions for VQA-RAD and 663 questions for SLAKE), whereas the answers of the open-form questions did not have any formatting restrictions (283 questions for VQA-RAD and 1,459 questions for SLAKE).

### 4.2 Metrics

We measured exact-match accuracy for the human-written questions that select answers from predefined candidates (all questions of CLEVR-Human and closed-form questions of VQA-RAD and SLAKE). For the open-form questions of VQA-RAD and SLAKE, we calculated  $LAVE_{GPT-4}$  [17] that uses an LLM (GPT-4 in this paper) for the evaluation of answers.

### 4.3 Benchmarks

#### 4.3.1 Benchmarks for CLEVR-Human

To benchmark the CLEVR-Human dataset, we employed three types of AI models, including general VLMs, VQA models trained with CLEVR (i.e., template-based VQA), and finetuned with CLEVR-Human (i.e., finetuned VQA). For the general VLMs, we utilized LLaVA [16] (version 1.5; 7B and 13B parameters; source code and model weights available on GitHub; see task instruction in Supplementary Figure 4) and GPT-4-Vision [1] (version 1106-preview; API used; see task instruction in Supplementary Figure 5), testing them on the CLEVR-Human validation dataset. For the template-based VQA models, we benchmarked FiLM [19], MAC [6], MDETR [9], and our CoQAH. In CoQAH, we combined MDETR trained with

**Table 1.** Comparison of accuracy between the general VLMs, template-based VQA models, and finetuned VQA models. CoQAH achieved the highest accuracy with large gaps compared to all the other general VLMs and template-based VQA models, although it did not utilize any data from CLEVR-Human for training. \* indicates accuracy measured using the CLEVR-Human validation data.

	Trained on CLEVR?	Finetuned on CLEVR-Human?	Acc. on CLEVR-Human (%)
<b>General VLMs</b>			
LLaVA (7B) [16]	No	No	43.8*
LLaVA (13B) [16]	No	No	47.6*
GPT-4-vision [1]	No	No	60.1*
<b>Template-Based VQA Models</b>			
FiLM [19]	Yes	No	56.6
MAC [6]	Yes	No	57.4
MDETR [9]	Yes	No	59.9 (59.5*)
<b>CoQAH</b>	Yes	No	<b>74.3*</b>
<b>Finetuned VQA Models</b>			
FiLM [19]	Yes	Yes	75.9
MAC [6]	Yes	Yes	81.5
MDETR [9]	Yes	Yes	81.7

**Table 2.** Comparison of the performance between the medical foundation and template-based VQA models on VQA-RAD and SLAKE. CoQAH reported the highest accuracy and  $LAVE_{GPT-4}$  in both open- and closed-form questions. There were huge gaps between the  $LAVE_{GPT-4}$  of the OFA-MIMIC and CoQAH, meaning that the chain of QA process in CoQAH indeed improved the correctness of the answers significantly.

	VQA-RAD		SLAKE	
	Acc. on Closed-Form (%)	$LAVE_{GPT-4}$ on Open-Form	Acc. on Closed-Form (%)	$LAVE_{GPT-4}$ on Open-Form
<b>Medical Foundation Models</b>				
BiomedGPT [27]	43.2	0.150	-	-
Med-Flamingo [18]	46.4	0.184	33.2	0.140
MedVInT-TD [28]	57.3	0.274	46.5	0.396
<b>Template-Based VQA Models</b>				
OFA-MIMIC [23]	59.5	0.095	69.4	0.097
<b>CoQAH</b>	<b>67.5</b>	<b>0.302</b>	<b>73.9</b>	<b>0.425</b>

CLEVR as a VQA model and GPT-4 as an LLM, measuring accuracy on the CLEVR-Human validation data. Finally, we compared the reported scores from the previous studies for the accuracy of the finetuned VQA models (FiLM, MAC, and MDETR).

#### 4.3.2 Benchmarks for VQA-RAD and SLAKE

To benchmark the VQA-RAD and SLAKE datasets, we employed three general VLMs specialized for the medical domain (i.e., medical foundation models): BiomedGPT [27] (no task instruction needed), Med-Flamingo [18] (see task instruction in Supplementary Figure 6), and MedVInT-TD [28] (see task instruction in Supplementary Figure 7). As template-based VQA models, we benchmarked OFA [23] trained by ourselves (i.e., OFA-MIMIC; using a model weight of  $OFA_{base}$  with 182M parameters; AdamW optimizer with learning rate =  $1e-4$ ,  $\beta_1 = 0.9$ , and  $\beta_2 = 0.999$ ; batch size = 16) because, so far, no VQA model trained with MIMIC-Diff-VQA is available to report performance on VQA-RAD and SLAKE. In CoQAH, we combined the OFA-MIMIC and GPT-4. Note that, different from the benchmarks on CLEVR-Human, we could not finetune the OFA-MIMIC using VQA-RAD or SLAKE because the number of chest X-rays is very limited (107 X-rays for VQA-RAD and 179 X-rays for SLAKE).

#### 4.3.3 Post-Processing of Answers

We observed that some of the answers the AI models provided frequently had the same meaning but different expressions (e.g., *radiograph* vs. *x-ray*, *pa views* vs. *pa*, *right side* vs. *right*, etc.), resulting in

misleading scores. Therefore, we substituted those answers as standard forms using post-processing (see Supplementary Table 3). We applied this post-processing to all the AI models we tested for fair comparison.

## 5 Results

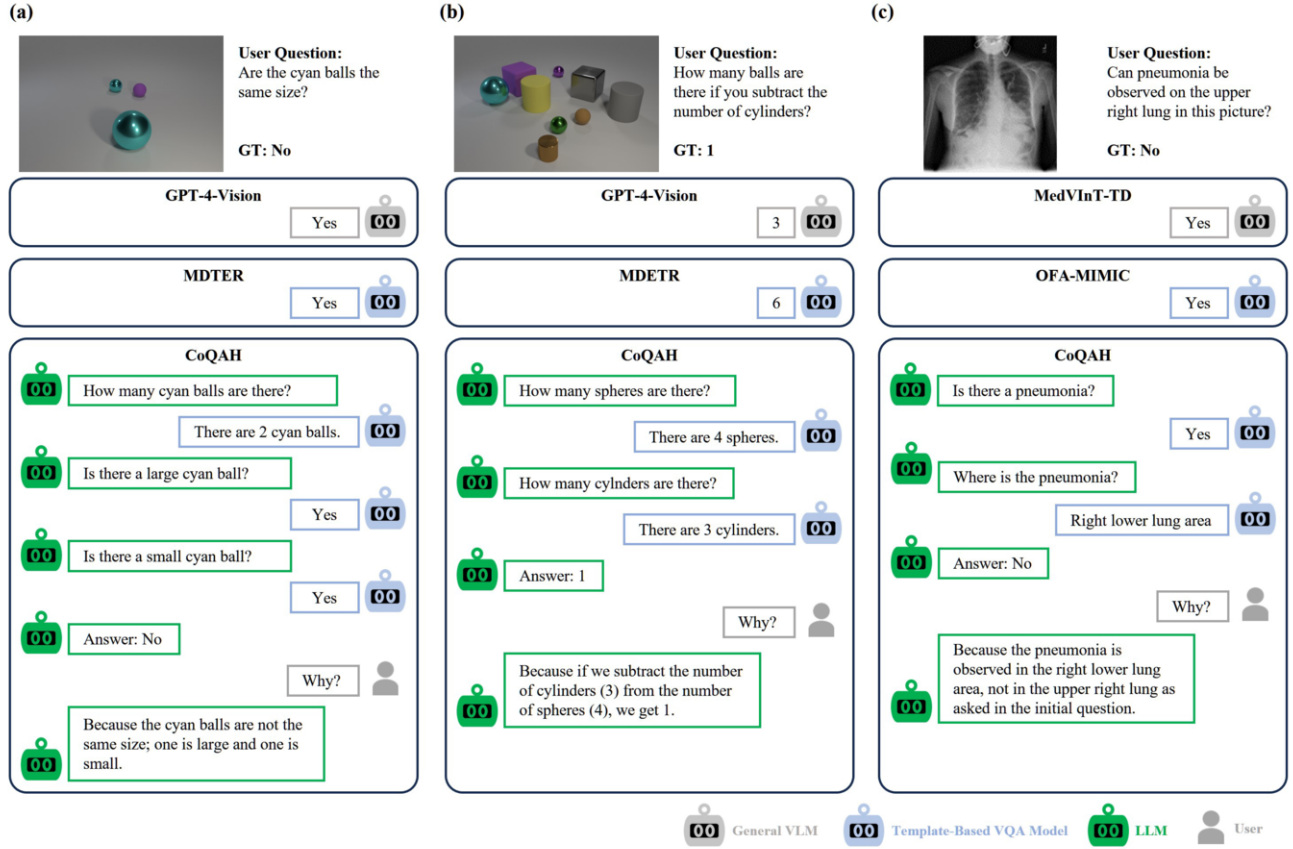
### 5.1 Benchmarks on CLEVR-Human

Table 1 compares accuracy between the general VLMs, template-based, and finetuned VQA models. Overall, the general VLMs reported lower accuracy than the template-based VQA models, except for the latest GPT-4-vision (e.g., 60.1% for GPT-4-vision vs. 59.9% for MDETR). CoQAH achieved the highest accuracy with large gaps compared to all the general VLMs and template-based VQA models (e.g., 74.3% accuracy for CoQAH vs. 59.9% for MDETR). There are still some performance gaps between CoQAH and the finetuned models (e.g., 74.3% for CoQAH vs. 81.7% for finetuned MDTER). However, it is worth noting that obtaining a large number of human-written QA pairs and finetuning a model is very difficult for most VQA tasks in general, and our framework substantially improves VQA performance using only synthetic QA pairs that are much easier to collect.

### 5.2 Benchmarks on VQA-RAD and SLAKE

Table 2 summarizes the performance of the medical foundation and template-based VQA models for VQA-RAD and SLAKE. Among all the models, CoQAH reported the highest accuracy in the closed-form





**Figure 4.** Example dialogues of CoQAH when we asked an LLM to interpret the reasons for the answers. In all cases, CoQAH successfully delineated the rationale behind the answers based on the information collected through the dialogues. For instance, for the question of *Are the cyan balls are the same size?*, CoQAH correctly explained the reason (*one is large and one is small*) for the response (*No*) based on the information from the two previous questions (*Is there a large cyan ball?* and *Is there a small cyan ball?*).

questions (e.g., VQA-RAD: 67.5% for CoQAH vs. 59.5% for OFA-MIMIC, SLAKE: 73.9% for CoQAH vs. 69.4% for OFA-MIMIC), and also the highest  $LAVE_{GPT-4}$  in the open-form questions (e.g., VQA-RAD: 0.302 for CoQAH vs. 0.274 for MedVInT-TD, SLAKE: 0.425 for CoQAH vs. 0.396 for MedVInT-TD). There were huge gaps between the  $LAVE_{GPT-4}$  of the OFA-MIMIC and CoQAH, meaning that the chain of QA process in CoQAH indeed improved the correctness of the answers significantly.

### 5.3 Interpreting Reasons for Answers

Since the CoQAH method utilizes an LLM to understand the context and finally answer the user question, we can ask the rationale of the final answer to the LLM (i.e., interpreting the AI model’s behavior), simply querying ‘why?’. For a few complex questions, we asked the LLM to explain the reasons to check whether it had logically driven their conclusions.

As shown in the examples of Figure 4, CoQAH successfully delineated the rationale behind the answers based on the information collected through the dialogues. For example, for the question of whether there are two cyan balls of the same size (Figure 4a), CoQAH reasonably explained the reason (*one is large and one is small*) for the answer (*No*) after checking the size of the cyan balls (*Is there*

*a large cyan ball?* and *Is there a small cyan ball?*). More examples can be found in Supplementary Figure 8 (for 3D-rendered images) and 9 (for chest X-rays).

### 5.4 Effect of the Maximum Number of Questions

The maximum number of questions to be queried is important parameter since it can directly affect the quality of the answers from CoQAH. Therefore, we measured the accuracy of CoQAH on the CLEVR-Human validation dataset (5% used), changing that number from 0 to 30. As shown in Table 3, The maximum accuracy was achieved when the maximum number of questions was 20, and even if we increased the number to 30, the accuracy was the same.

### 5.5 Ablation Study

We conducted an ablation study by excluding several components in CoQAH and measured the accuracy based on the CLEVR-Human validation data (5% used). The results of the ablation study are summarized in Table 4. When we ablated EUH in CoQAH, the accuracy on CLEVR-Human was slightly decreased (75.8% for CoQAH vs. 74.7% without EUH). When ablating the step-by-step QA process in

**Table 3.** Accuracy on CLEVR-Human validation dataset according to the changes of the maximum number of questions to be queried in CoQAH. We only used 5% of the validation data for the ablation study. Until the maximum number of questions reached 20, accuracy increased as the number of questions increased. However, the accuracy remained the same when the maximum number of questions was set at 20 and 30.

Maximum number of questions to be queried	0	5	10	20	30
Acc.on CLEVR-Human (%)	60.0	68.1	69.4	75.8	75.8

**Table 4.** Comparison of accuracy on CLEVR-Human by ablating the step-by-step QA process, EUH, or the template format description. Ablating the QA process degraded the accuracy significantly. Ablating the format description dramatically degraded the accuracy, meaning that the questions from the LLM must conform to the template format.

	Acc. on CLEVR-Human (%)
CoQAH	75.8
without EUH	74.7
without step-by-step QA process	63.1
without format description	35.6

CoQAH by enforcing the LLM to ask all the questions simultaneously (*ask me up to 20 questions all at once*), the accuracy was degraded mainly (75.8% for CoQAH vs. 63.1% without the process), indicating the importance of this process to derive the answers correctly. Finally, when we ablated the format description in the task instruction by letting LLM generate any form of questions during the chain of QA process, the accuracy was dramatically degraded (75.8% for CoQAH vs. 35.8% without the format description). This means that the questions from the LLM must conform to the template format so the VQA model can understand those questions. Additionally, to investigate the effect of few-shot prompting on CoQAH, we compared the performance of CoQAH in zero-shot, one-shot, and two-shot settings. Dialogues between the VQA model and the LLM were provided as few-shot exemplars to the LLM. The results are shown in Table 5. The performance increased when few-shot exemplars were given, but increasing the number of exemplars from one to two did not further improve performance.

### 5.6 Error Analysis of CoQAH

We performed an error analysis by investigating the dialogues where CoQAH failed to answer correctly (5% CLEVR-Human validation used) and categorized them according to the reasons (e.g., incorrect answers from VQA model). The results of the error analysis are shown in Table 6.

The most significant portion of errors (48%) occurred when the LLM failed to query key questions or made a hasty conclusion after asking some irrelevant questions (e.g., Supplementary Figure 8e). It means that the LLM still has shown a limited reasoning ability during the QA process for some complex questions. The second largest portion (25%) was due to incorrect responses from the VQA model during the QA process (e.g., Supplementary Figure 8d). Some errors (11%) dedicated to the questions where the LLM was impossible to answer by querying template-based questions only (e.g., Supplementary Figure 8f). Other errors (15%) included failures of following instructions, data format (e.g., *The answer should be in {}*), and so on.

**Table 5.** The performance of CoQAH in zero-shot, one-shot, and two-shot settings. Dialogues between the VQA model and the LLM were provided as few-shot exemplars to the LLM. Performance increased when few-shot exemplars were provided, but adding more exemplars, from one to two, did not result in any further improvement.

Prompt Setting	Zero-shot	One-shot	Two-shot
Acc. on CLEVR-Human (%)	75.8	80.0	79.4

**Table 6.** Results of the error analysis for CoQAH. We investigated the dialogues where CoQAH failed to answer correctly and categorized them according to the reasons. The most significant portion of errors occurred when the LLM failed to query key questions or made a hasty conclusion after asking some irrelevant questions, showing a limited reasoning ability.

Error Category	Percentage (%)
Limited reasoning ability of LLM	48
Incorrect response from VQA model	25
Impossible to answer by querying template-based questions only	11
Other errors	15

## 6 Conclusion

In this study, we proposed a novel CoQAH methodology that answers human-written questions via a reasoning process based on a dialogue. Throughout the extensive experiments, we have shown that CoQAH achieved state-of-the-art performance on various datasets of different images (3D-rendered and chest X-ray) without finetuning using the human-written questions. We believe that this new method potentially promotes the wide applications of AI for visual question answering, improving the practicality of AI in real environment.

## 7 Limitation

First, we could only test a limited number of general VLM models (LLaVA [16] and GPT-4-Vision [1]) for CLEVR-Human since, currently, many LLMs do not support taking an image as an input. Second, we reported the accuracy of the AI models we ran on CLEVR-Human validation data because the test set was unavailable. Third, even though we tried our best to optimize the task instructions for the general VLM and medical foundation models, those might not be the best due to the lack of methods to tune the instructions. Fourth, we only validated CoQAH using two types of images (3D-rendered and chest X-ray) because of a limited number of template-based datasets.

## References

- [1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [2] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.
- [3] J. P. Bigham, C. Jayant, H. Ji, G. Little, A. Miller, R. C. Miller, R. Miller, A. Tatarowicz, B. White, S. White, et al. Vizviz: nearly real-time answers to visual questions. In *Proceedings of the 23rd annual ACM symposium on User interface software and technology*, pages 333–342, 2010.
- [4] J. Guo, J. Li, D. Li, A. M. H. Tiong, B. Li, D. Tao, and S. Hoi. From images to textual prompts: Zero-shot visual question answering with

- frozen large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10867–10877, 2023.
- [5] X. Hu, L. Gu, Q. An, M. Zhang, L. Liu, K. Kobayashi, T. Harada, R. M. Summers, and Y. Zhu. Expert knowledge-aware image difference graph representation learning for difference-aware medical visual question answering. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4156–4165, 2023.
  - [6] D. A. Hudson and C. D. Manning. Compositional attention networks for machine reasoning. *arXiv preprint arXiv:1803.03067*, 2018.
  - [7] J. Johnson, B. Hariharan, L. Van Der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017.
  - [8] J. Johnson, B. Hariharan, L. Van Der Maaten, J. Hoffman, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick. Inferring and executing programs for visual reasoning. In *Proceedings of the IEEE international conference on computer vision*, pages 2989–2998, 2017.
  - [9] A. Kamath, M. Singh, Y. LeCun, G. Synnaeve, I. Misra, and N. Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021.
  - [10] T. Kim. Source code for "generalizing visual question answering from synthetic to human-written questions via a chain of QA with a large language model", 2024. Available at <https://github.com/tae2hee/CoQAH>.
  - [11] T. Kim, Y. Cho, H. Shin, Y. Jo, and D. Shin. Generalizing visual question answering from synthetic to human-written questions via a chain of QA with a large language model. *arXiv preprint arXiv:2401.06400*, 2024. Full version of this paper.
  - [12] J. J. Lau, S. Gayen, A. Ben Abacha, and D. Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018.
  - [13] Y. Li, Y. Liu, Z. Wang, X. Liang, L. Liu, L. Wang, L. Cui, Z. Tu, L. Wang, and L. Zhou. A comprehensive study of gpt-4v's multimodal capabilities in medical imaging. *medRxiv*, pages 2023–11, 2023.
  - [14] Z. Lin, D. Zhang, Q. Tao, D. Shi, G. Haffari, Q. Wu, M. He, and Z. Ge. Medical visual question answering: A survey. *Artificial Intelligence in Medicine*, page 102611, 2023.
  - [15] B. Liu, L.-M. Zhan, L. Xu, L. Ma, Y. Yang, and X.-M. Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE, 2021.
  - [16] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.
  - [17] O. Mañas, B. Krojer, and A. Agrawal. Improving automatic vqa evaluation using large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4171–4179, 2024.
  - [18] M. Moor, Q. Huang, S. Wu, M. Yasunaga, Y. Dalmia, J. Leskovec, C. Zakkai, E. P. Reis, and P. Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR, 2023.
  - [19] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.
  - [20] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
  - [21] A. M. H. Tiong, J. Li, B. Li, S. Savarese, and S. C. Hoi. Plug-and-play vqa: Zero-shot vqa by conjoining large pretrained models with zero training. *arXiv preprint arXiv:2210.08773*, 2022.
  - [22] M. Tsimpoukelli, J. L. Menick, S. Cabi, S. Eslami, O. Vinyals, and F. Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021.
  - [23] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, and H. Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022.
  - [24] Z. Yang, Z. Gan, J. Wang, X. Hu, Y. Lu, Z. Liu, and L. Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3081–3089, 2022.
  - [25] H. You, R. Sun, Z. Wang, L. Chen, G. Wang, H. Ayyubi, K.-W. Chang, and S.-F. Chang. Idealgpt: Iteratively decomposing vision and language reasoning via large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11289–11303, 2023.
  - [26] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.
  - [27] K. Zhang, J. Yu, Z. Yan, Y. Liu, E. Adhikarla, S. Fu, X. Chen, C. Chen, Y. Zhou, X. Li, et al. Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. *arXiv preprint arXiv:2305.17100*, 2023.
  - [28] X. Zhang, C. Wu, Z. Zhao, W. Lin, Y. Zhang, Y. Wang, and W. Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023.
  - [29] K. Zhou, K. Lee, T. Misu, and X. E. Wang. Vicor: Bridging visual understanding and commonsense reasoning with large language models. *arXiv preprint arXiv:2310.05872*, 2023.