b
UNIVERSITÄT
BERN

Graduate School for Cellular and Biomedical Sciences
UNIVERSITY OF BERN

Spatial Awareness and Logic for Robust Visual Question Answering

PhD Thesis submitted by

Tascón Morales Sergio

for the degree of PhD in Biomedical Engineering

Supervisors
Prof. Dr. Raphael Sznitman
Dr. Pablo Márquez Neila
Faculty of Medicine of the University of Bern

Co-advisor
Dr. Damien Teney
Idiap Research Institute, Martigny, Switzerland

Accepted by the Faculty of Medicine, the Faculty of Science and the Vetsuisse Faculty of the University of Bern at the request of the Graduate School for Cellular and Biomedical Sciences

Bern, Dean of the Faculty of Medicine

Bern, Dean of the Faculty of Science

Bern, Dean of the Vetsuisse Faculty Bern

UNIVERSITY OF BERN
Graduate School for Cellular and Biomedical Sciences
Faculty of Medicine

Abstract

PhD in Biomedical Engineering

Spatial Awareness and Logic for Robust Visual Question Answering

by Tascón Morales Sergio

In recent years, deep learning models have become an integral part of the daily lives of millions, extending their influence into specific domains such as medicine. The integration of vision and language capabilities has notably facilitated smoother interactions between users and models. Questions and answers have long served not only as a means of interaction with machines but also as a test for evaluating their level of intelligence. In particular, inquiries related to visual content, encapsulated by Visual Question Answering (VQA), provide a mechanism to probe a model's visual understanding. In the medical domain, this aspect holds considerable significance, given the crucial role that trust plays in the adoption of these systems by medical professionals. However, the often opaque nature of most models hinders the assessment of true visual understanding, concealing potential shortcuts and biases. Crucial aspects of reasoning, such as compositionality and consistency, are at times overlooked in favor of high overall performance. In line with this perspective, this work introduces several contributions in the domains of localized questions and consistency for VQA.

The first part of the thesis explores questions about specific image regions. Two distinct methodologies are proposed. The first method employs a localized attention mechanism, integrating information about the target region through a binary mask. Localized attention allows the network to consider contextual cues necessary for answering the question, focusing subsequently on the region specified by the user. The second method extends the concept of localized questions to Multimodal Large Language Models (MLLMs) by introducing targeted visual prompting. Here, a customized visual prompt is formulated, encompassing the isolated region and its contextual representation within the image.

The second part of the thesis focuses on avoiding contradictions by enhancing consistency. The first method involves categorizing queries as perception vs. reasoning questions and utilizing a loss function term to penalize inconsistencies during training. The second method proposes a broader interpretation of consistency in VQA based on logical relations and introduces an auxiliary method for predicting these relations. Similar to the first method, this approach employs a loss term to enforce more consistent behavior during the training phase.

Acknowledgements

Every book needs an author. My life contains pages of many types; some are blurry, some are crumpled, some are torn, some are black and white, some are colorful, some are stunning, some are full, and some are empty. I can only be thankful to my Author when I reread them. Though the pages looked like a randomly assembled collage at some point, they make sense together today. Through His mercy and love, I trust that the empty pages will be filled with everlasting words and art.

Every pupil needs a master, a role model, someone who started the journey first and is willing to guide, share knowledge, encourage, and correct. I was fortunate enough to have two of them during this adventure. I thank Prof. Dr. Raphael Sznitman for granting me the opportunity to work in his lab under his supervision. For looking at me in full size and not as a dot from the humongous mountain he has climbed. For being humble and kind, for always being willing to help and to discuss, and for patience. I thank Dr. Pablo Márquez-Neila for not letting me swallow ideas without chewing on them first and for always being alert at details that often pointed out my errors. I also thank him for his humility and care for others' processes and deadlines. I thank both for teaching me valuable lessons about research, communication, and life.

Every explorer needs a compass to navigate unknown paths. This was especially true during this research journey. I thank my co-advisor, Dr. Damien Teney, for his interest in my process and his willingness to share his perspectives in his field of expertise. His ideas and work inspired me to pursue more profound knowledge and realize the limits of conventional ideas. I thank my mentor, PD Dr. Sigve Haug, for assuming his role with the best disposition. I also owe my thanks to the University of Bern and the Swiss National Science Foundation (SNSF) for their relevant support in funding my research.

Every combatant needs a squad. At the AIMI lab, I found a team of brilliant and competent people who selflessly offered me their friendship and support. First, I thank those who, at the beginning, made things easier for me and offered to help, including Vasily Tolkachev, Dr. Mathias Gallardo, Dr. Tatiana Fountoukidou, and Dr. Thomas Kurmann. I thank the other Ph.D. students, Javier Gamazo-Tejero, Lars Doorenbos, Lukas Zbinden, Fei Wu, Paulo Sampaio, and the later-arrived Moritz Schmid. The moments we shared made the process fun and interesting, and I certainly learned something from each of them. My gratitude also goes to all the other members of the team, including Prisca Dotti, Dr. Negin Ghamsarian, Marta Colmenar Herrera, Dr. Alain Jungo, Dr. Theodoros Pissas, Dr. Christopher Hahne, Dr. Corin Otesteanu, Michel Hayoz, Davide Scandella, Tamara Danilovska, and everyone else I had the pleasure of sharing with, even if for a short time. Finally, I thank the administrative staff, Pia Eichholzer, Diana Bethge, and Daniela Schmidt, for always being willing to help.

I was lucky to have an extended squad of people who supported me and shared with me in person or from thousands of kilometers away. I thank Jose Arturo García, William Giraldo, Sergio Castro, Ivan Said Góngora, Eduardo Osorio, Miguel García, Nicolás Camargo, Cassandra Chetty and her lovely family, Andreina Ravani, Denis Borel, Emre Hacan and his wife Gülcan,

Yasemin Aversa, Didem Ekrem, Gresa Asani, and my MAIA friends, specially Roa'a Khaled, Valeria Abramova, Tewodros Weldebirhan Arega and Ahmed Gouda. I also extend my thanks to Umut and Gökcen Parlar and their parents, Canan and Turgay, for their kindness, support, and positive attitude.

Every tree needs roots. I call my roots family, the people who nurture me and are the foundation of the person I am today. I want to thank my family, including my parents, Cruz Stella and Gustavo, my siblings Natalia and Juan Felipe, and my little niece, Violeta. Their love, prayers, sacrifices, and constant support have been a blessing that distance cannot extinguish. I also thank my aunt Sandra and her husband Pepe, my aunts Libia, Liliana, Neyla, and Cecilia, my uncle Humberto, my cousins Sandra Teresa, Monica, Edward, Iván, and Daniela, and my grandma Teresa.

Every heart needs love, someone who is always there, no matter what, who confronts your deepest shadows and is not terrified, who sees in you what other people cannot see. I met Ebru Parlar more than a thousand days ago, and since the beginning, her love, company, support, understanding, care, and cuteness have been priceless treasures in my life. I thank her for believing in me, helping me grow, and being a shiny little star that illuminates my days.

Contents

List of Figures

List of Tables

List of Abbreviations

2D two-dimensional 4

3D three-dimensional 4

AI Artificial Intelligence 6, 12

ANN Artificial Neural Network 3

BAN Bilinear Attention Networks 28, 76, 77

BERT Bidirectional Encoder Representations from Transformers 20, 23, 71, 74, 76, 81

BoW Bag of Words 13

BRNN Bidirectional Recurrent Neural Network 14, 15

BUTD Bottom-Up Top-Down 27

CNN Convolutional Neural Network 4, 7, 9, 11, 22, 27, 28, 93, 95

DAE Denoising Auto-Encoder 29

DME Diabetic Macular Edema 9, 30, 31, 39, 49, 57, 58, 61, 62, 66, 75–78

GELU Gaussian Error Linear Unit 22

GRU Gated Recurrent Unit 16

ICL In-Context Learning 25

LLM Large Language Model 5, 7, 9, 12, 19–22, 24, 25, 28, 45–48, 50, 53, 93–95

LSTM Long Short-Term Memory 12, 15, 16, 27, 37, 38, 40, 63

MCB Multimodal Compact Bilinear 27, 28

Med-VQA Medical Visual Question Answering 7, 8, 25, 26, 28–30, 36, 45, 46, 57, 58, 70, 85, 90, 91, 93, 94

MFB Multimodal Factorized Bilinear 27, 28

MFH Multimodal Factorized High-order 27

ML Machine Learning 3

MLB Multimodal Low-rank Bilinear 27, 28

MLLM Multimodal Large Language Model 9, 24, 28, 29, 45–47, 50, 86, 90, 93, 96

MUTAN Multimodal Tucker Fusion for Visual Question Answering 27

NLG Natural Language Generation 12

NLI Natural Language Inference 10, 70, 75, 81

NLP Natural Language Processing 9, 11, 12

NLU Natural Language Understanding 12

PEFT Parameter-Efficient Finetuning 22

R-CNN Region-based Convolutional Neural Network 28

ReLU Rectified Linear Unit 22, 41

RNN Recurrent Neural Network 5, 7, 11–16, 93, 94

ROC Receiver Operating Characteristic 40, 81

RTE Recognizing Textual Entailment 70

SAN Stacked Attention Networks 27, 28

ViT Vision Transformer 4, 9, 11, 22, 23, 50, 95

VLM Vision-Language Model 9, 11, 24, 69, 96

VQA Visual Question Answering 6–11, 25–30, 35–37, 39, 40, 43, 46, 47, 50, 53, 57–61, 63, 65–79, 81, 86–88, 90, 93–97

VTT Visual Turing Test 7

For my mom Cruz Stella and my sister Natalia,
whose infinite love constantly rekindles my heart.

For Ebru, whose smile turns days into miracles,
whose eyes can quell the fiercest storms.

1 Introduction

Computers, in various forms, have become an integral aspect of human daily life. Their ability to tackle a diverse range of tasks, coupled with their efficiency, has significantly expanded their applicability across various fields in recent years. Notably, in the field of medicine, computers have played and continue to play a crucial role, leveraging their potential to assist medical experts in the analysis and annotation of medical data. The capacity to comprehend both textual and visual information is a pivotal feature central to models that can assist medical experts in their daily tasks.

The interaction with medical images by means of written questions holds particular importance, as it facilitates an evaluation of the machine's actual understanding of the information and its ability to reason effectively to provide accurate answers. Within this introductory chapter, we delve into the breakthroughs and concepts that have paved the way for the extensive capabilities computers offer broadly in diverse scenarios and specifically in the medical domain. This exploration is done drawing from the essence of Alan Turing's groundbreaking work about intelligent machines.

1.1 Reading Words, Seeing Worlds and Asking Questions

1.1.1 Thinking Machines

The history of devices capable of aiding in computational tasks extends back at least four millennia, beginning with the creation of the abacus [1]. Evolving from this rudimentary counting tool, subsequent centuries revealed more intricate ancient mechanical devices, such as the Antikythera mechanism [2] and the astrolabe [3], originating from ancient Greece and utilized for astronomical purposes. In more recent history, the seventeenth-century introduction of the slide rule represented a significant step toward more efficient mathematical operations [4]. While these devices proved useful, their design centered around task-specific manipulation, where the instructions they executed were pre-defined either within the device or by the operator at the time of execution.

Charles Babbage, acknowledged as the father of the computer, introduced a more flexible computing system in the early nineteenth century. His mechanical computer was *programmable*, allowing for the sequential execution of an ordered collection of instructions (*i.e., a program*) defined by the user for a specific task. The program, along with any input data, was provided to the device using punched cards [5]. This concept of programmable computers persisted, but the implementation transitioned from mechanical operation to vacuum tubes and subsequently to transistors. At the time, the first electronic computers were large and heavy devices that only some institutions had the privilege to utilize.

In 1950, Alan Turing published a work titled Computing Machinery and Intelligence [6], where he addressed the question “Can machines think?" by framing it as the outcome of a game. The game, known as the imitation game or the Turing test, involves three participants in isolation, as illustrated in Fig. 1.1: a machine (A), a person (B) and another person assuming the role of interrogator (C). In this scenario, the interrogator uses text-based communication to interact with A and B through questions and answers. The goal for A is to behave in a manner indistinguishable from a human during the conversation. Following the game, interrogator C indicates which participant corresponds to the computer and which is human. From this perspective, if C incorrectly classifies the participants with high probability, the machine is considered intelligent or capable of thinking.

While specific practical and philosophical limitations in the imitation game have been identified [6, 7], it underscores the importance of language understanding and generation in machines. Moreover, it highlights the key role of questions and answers in evaluating the true intelligence of a machine. Additionally, beyond merely determining a machine’s intelligence, the use of questions and answers represents a means of communication for the execution of specific tasks. An intelligent machine, as defined by the Turing test, not only communicates like a human but also possesses the capability to perform tasks akin to human abilities, rendering it versatile across a broad spectrum of applications, some of them in the field of medicine.

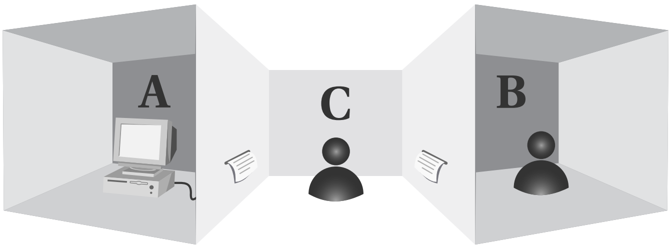


FIGURE 1.1: Illustration of the imitation game. A computer (A) and a person (B) answer questions posed by a human interrogator (C). Participant A aims at providing human-like answers, attempting to trick C into thinking that the answers were provided by a person.

In order to emulate human behavior, the machine is expected to learn. While Turing explored some intriguing ideas such as the use of rewards and punishment in the learning process, it was not until the advent of Machine Learning (ML) that machines began to perform tasks with a degree of acquired knowledge (*i.e.*, *learning*). This initiation occurred with the work of Arthur Samuel in the 1950s, where he proposed a learning method for the game of Checkers based on the optimization of a game tree [8]. A major breakthrough occurred with the perceptron [9], a single-layer neural network with a linear threshold function, considered to be an essential building block of modern Artificial Neural Networks (ANNs). The perceptron updated its parameters using the difference between the output of the target. Stacking multiple perceptrons required the propagation of errors through the network, which was enabled with backpropagation [10]. This advancement facilitated the training of multilayer perceptrons [11, 12] and became the standard algorithm for error propagation. Together with gradient descent, it allows models to learn from experience. This stands in contrast to the traditional approach of programming models with pre-defined instructions for every conceivable input and state.

Returning to Turing's work, the aforementioned developments paved the way for machines to resemble the way in which humans learn from experience more closely. The process of learning to perform specific tasks, such as playing Checkers or Chess [13], marked only the beginning of Turing's notions into the realm of machine intelligence. As he articulated in his paper,

> *It can also be maintained that it is best to provide the machine with the best sense organs that money can buy, and then teach it to understand and speak English.*

We now delve into the role of vision and language understanding in enabling the interrogation of a machine.

1.1.2 Computer Vision and Language Processing

Perceiving the World

Providing machines with vision capabilities marked a significant stride toward realizing the machine intelligence envisioned by Turing. However, it is crucial to distinguish between image processing and computer vision. Image processing aims to enhance visual appearance, extract information, or transform images. On the other hand, computer vision is focused on recovering the 3D structure of the world from input images and utilizing this information for full scene understanding [14]. In essence, computer vision is concerned with understanding the reality expressed by an image or video.

Early computer vision approaches were introduced in the 1970s with the goal of achieving comprehensive scene understanding. These approaches encompassed a diverse range of techniques, including 3D structure inference from 2D lines, line labeling, generalized cylinders, pictorial structures, and optical flow algorithms [14]. In 1980, a network architecture named *Neocognitron* emerged and was applied to handwritten character recognition, building upon earlier ideas derived from the visual nervous system. The concept involved cascading a series of alternating simple and complex cells, where the former was responsible for local feature detection, and the latter captured global patterns [15]. This pioneering work laid the groundwork for Convolutional Neural Networks (CNNs) [16], which share structural similarities with Neocognitron but are more flexible due to the use of generic convolution and pooling operations, facilitating the automatic extraction of hierarchical feature representations. Despite being formalized at the end of the twentieth century, the widespread application of these networks was hindered by hardware limitations [17].

Starting around 2012, as a result of the developments in graphics processing units and parallelization, the utilization of CNNs experienced exponential growth, witnessing the proposal of various architectures tailored for tasks like image classification, image segmentation, facial recognition, and image generation [18]. More recently, an alternative architecture has emerged as a formidable contender to CNNs. The Vision Transformer (ViT) [19] adapts an architecture originally designed for text processing to operate effectively with images. Both CNNs and ViTs can be regarded as the closest approximation to the “sense organs" envisioned by Turing. Attaining meaningful representations or descriptors of images represents a crucial step toward machines capable of performing tasks akin to human abilities. These architectures have also made substantial inroads into the medical domain, where they can, for example, contribute to alleviating the workload of clinical experts who may struggle to analyze all available images or benefit from a second opinion provided by a precise automated system.

However, models focused solely on vision are generally trained for specific tasks and often offer limited interaction, if any, with users. As Turing highlighted, imparting machines with the capability to understand and communicate in natural language is a crucial step forward.

Reading and Saying Words

The desire for machines capable of understanding language is not a recent aspiration for humanity. In the mid-1930s, patents were conceived, with specific instances of translation machines [20]. While rudimentary and basic, these can be considered the earliest practical ideas attempting language processing for a specific task. During the 1940s and 1950s, the challenge of machine language translation spurred initial research efforts, primarily following a rule-based design. Given the limited access to computers during this era, primarily research and military institutions could engage in such endeavors. Consequently and due to political interests, the main objective was to create machines capable of translating Russian text into English [21]. Linguistic experts like Noam Chomsky identified limitations in translation systems during the 1950s, highlighting a lack of genuine comprehension of text content [22].

Gradually, the applications of language transitioned from translation to dialogue, bringing the process a step closer to the imitation game, albeit with limited results. An illustration of this shift is ELIZA [23], which engaged in simulated conversations with humans by utilizing reassembly and decomposition rules. This program, however, lacked a genuine understanding of the meaning of words.

Years later, in 1970, as a response to the disappointment stemming from earlier works and with an aim to incorporate the meaning of words along with syntactical analysis and identification of lexical items, a program named SHRDLU was introduced [24]. This program demonstrated the capability to answer basic questions, execute commands, and augment its knowledge about a simulated robot arm with access to toy objects. Subsequent systems further expanded the ability to answer questions to specific fields, such as lunar geology [25].

An important advancement in the representation of sequential information came about with Recurrent Neural Networks (RNNs) [12]. When coupled with the error propagation and parameter adjustment techniques mentioned earlier, RNNs played a crucial role in enhancing the extraction of meaningful information from sequences and the sequential generation of data. The concept of mapping words to distributed vectorial representations that consider similarity [26] (*i.e.*, word embeddings) marked a significant step, enabling the application of RNNs, in all of its variations, to text sequences for various tasks.

More recently, the transformer architecture [27] has revolutionized language processing by addressing limitations of RNNs such as long-term dependencies, scalability and efficiency. This architecture, combined with substantial computing power and large text datasets, has paved the way for the conception and implementation of Large Language Models (LLMs). LLMs are language models with a large number of parameters and trained on internet-scale datasets. In recent years, these models have evolved to the point of becoming prominent as information sources, chatbots, or assistants [28].

Having models that perceive the world and “speak" a language seems to align with Turing’s aspirations for the imitation game. However, devising a model that seamlessly integrates vision and language is not a trivial task, as it demands a certain level of correspondence between text and vision representations.

1.1.3 Visual Questioning

Endowed with vision and language capabilities, machines can perform a diverse array of tasks, such as image captioning, image retrieval, text-to-image generation, visual grounding, visual storytelling and Visual Question Answering (VQA). Notably, VQA holds considerable appeal as it enables interaction between humans and machines through questions related to images or videos. Involving skills like spatial reasoning, logical inference, comparisons, counting, memorization, object and attribute recognition and transitive relation tracking, VQA requires visual understanding at various levels [29]. This enhanced visual understanding adds a dimension to the Turing test, allowing the assessment of a model's ability to comprehend and interpret visual information. In this scenario, the setup for the Turing test is expanded: Interrogator C poses questions about visual content, which participants A and B can perceive, and they provide answers. Similar to the standard Turing test, if the interrogator cannot reliably distinguish between human and computer responses, the machine is deemed to pass the test.

Early applications towards visual question systems can be traced back to 2003, where applications included real-time motion tracking for browsing surveillance videos [30] and questions about news videos based on analysis of transcripts but informed by computer vision techniques [31]. These systems, however, had inherent limitations, such as a restricted set of possible questions and the reliance on an external module for vision processing, neglecting the necessity for a joint understanding of both modalities.

A few years later, the concept of *Photo-based Question Answering* emerged as a more comprehensive task involving questions about objects within in an image [32]. As depicted in Fig. 1.2, the system comprises three layers: the first layer matches the input image to web images and extracts structured data from multimedia databases, the second layer searches for appropriate answers in an internal repository, and the third layer delegates more complex questions to humans.

A real-world application of visual questions emerged in 2010 with VizWiz [33], designed to provide visually impaired individuals with answers to questions about their daily interactions with the environment. Initially relying on crowd-sourced workers, the platform evolved to incorporate Artificial Intelligence (AI) solutions in subsequent years [34, 35]. Notably, this application underscores the importance of free-form questions, allowing users to employ any grammatical structure to inquire about the contents of an image.

Further developments broadened the landscape in terms of datasets and methodologies. One approach enhanced the surveillance video browsing application mentioned earlier by utilizing a probabilistic model to capture relations between video and text using parse graphs [36].

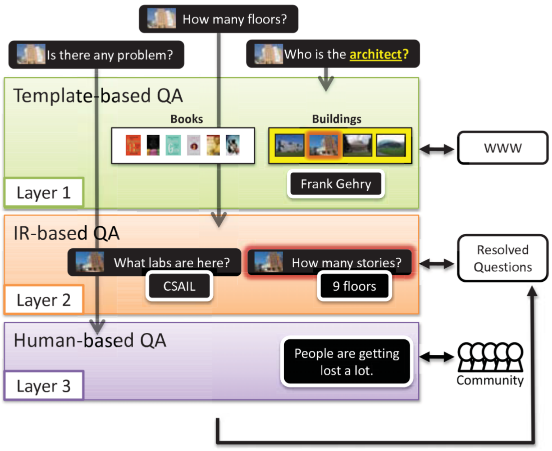


FIGURE 1.2: A three-layer method for photo-QA. The first layer performs image matching of the input image to web images and extracts structured data. The second layer searches for applicable answers. The third layer allows humans to answer more complicated questions. From [32].

Another approach employed segmentation to gather facts about objects for answering template-generated questions from a limited vocabulary [37]. Yet another formalized the concept of Visual Turing Test (VTT) to evaluate the visual understanding of machines through a sequence of binary questions, ensuring that the history of questions and correct answers was unhelpful in answering the current question [38].

The formal pursuit of answering visual questions gained momentum in 2015 with the introduction of the VQA task. This challenge featured a dataset with thousands of open-ended human-generated questions about images [39] and presented an architecture as a baseline for benchmarking. The architecture comprised a frozen CNN for image encoding, an RNN for question encoding, a multiplication operation to combine both embeddings and an output classifier to select the most likely answer from a predefined list. Over time, the type of answer generated by models has evolved, with LLMs enabling the generation of more detailed answers and descriptions, aligning with Turing's vision of machines generating human-like responses.

Following the introduction of VQA for natural images, the task found its way into the medical domain, garnering attention and inspiring researchers. Advancements in Medical Visual Question Answering (Med-VQA) have closely mirrored those in classical VQA, with some exceptions for addressing data-related challenges and specialized architectures. A more detailed history of VQA architectures and datasets is offered in Sec. 2.4.2.

Expanding on this trajectory, we can envision a variation of the Turing test for medical images. In this scenario, participants B and C are replaced by experts in a specific medical imaging modality, with A being the machine. Interrogator C poses medical questions about images to A and B. If C tends to believe that A is a medical expert with high probability, the machine could be deemed intelligent in a medical sense, showcasing specialized knowledge beyond general human knowledge. In this context, the accuracy of answers and the terminology used play a crucial role, demanding more from the machine to simulate a medical expert compared to simulating a human. One case in which this simulation can fail is when the machine provides contradictory answers too often. Following this line of thought, we briefly examine the case in which two questions are asked about the same image.

1.2 Making Sense

Given that humans are prone to errors, it is reasonable to expect a machine emulating human behavior to also make mistakes, especially when faced with challenging tasks that allow only limited generalization to unseen examples post-training, introducing errors in responses. However, in the imitation game, the nature of errors made by A and B can significantly impact C's final identification of the participants as machine or human.

Illustrating this with the text-only Turing test, consider the following example. If we query the machine about the years Abraham Lincoln was alive and receive the correct response "1809 to 1865," but then ask about the century and get the incorrect answer "18th century," we identify an issue beyond mere errors. Abraham Lincoln being alive both in 1809 - 1865 and in the 18th century is a contradiction. Asking about the same information, we expect a machine (as we do a human) to avoid contradictions in responses, displaying *consistency*.

Detecting such contradictions, a skill innate to humans, proves challenging for machines but directly influences the quality of reasoning they employ [40]. Reasoning, involving “scaling to ever-larger search spaces and understand the world broadly," implies consistency, causality, and compositionality [41]. The absence of any of these elements can cast doubt on the quality of reasoning.

Incorporating images into the imitation game facilitates testing the model's consistency, as queries can reference external visual evidence. In the text-only scenario, a comprehensive image description (objects, relations, color, structure, etc.) would be needed in the question, creating a challenge. For instance, consider presenting a VQA model with an image of a bear statue and asking: "What is this?" and "Is it alive?" If the model responds "a statue of a bear" and "yes," respectively, inconsistent behavior becomes apparent. Humans leverage logic and prior knowledge, understanding that a statue cannot be alive. Thus, ensuring machine consistency requires integrating logical faculties, explicitly or implicitly.

In the medical domain, the significance of consistent answers amplifies due to the potential impact on medical decisions. The adoption of Med-VQA systems by medical experts hinges on trust, with models demonstrating less contradictory behavior being perceived as more trustworthy and effective tools.

1.3 Thesis Statement

This thesis addresses visual understanding and reasoning in VQA by means of two perspectives:

1. Localized queries, where questions can be posed about any region of an image,
2. Consistency enhancement, where a model is encouraged to avoid contradictions,

respectively. Considering this, we formulate the following thesis statement:

Achieving high-quality clinical decisions through Visual Question Answering systems requires a prioritization of consistency and fine-grained queries, offering a pathway to improved spatial understanding and overall model reliability.

1.4 Organization and Contributions of the Thesis

Chapter 2 lays the foundation with key concepts related to language and vision, along with their combination. The chapter delves into Natural Language Processing (NLP), highlighting its prominent architectures, and focuses on pivotal aspects of computer vision, emphasizing CNNs and ViTs. Vision-Language Models (VLMs) are then explored, followed by an in-depth examination of VQA from both architectural and data perspectives. Additionally, basic concepts regarding Diabetic Macular Edema (DME) staging are presented, due to their relevance in this thesis.

Part I is dedicated to the exploration of localized questions (*i.e.*, questions about specific image regions) in VQA. Chapter 3 introduces a method enabling such questions for VQA models with guided attention mechanisms. The proposed approach involves localized attention, integrating a target region represented by a binary mask into the VQA's attention mechanism. This enables the model to compute attention maps on the entire image, subsequently filtering them spatially to focus on the specified region. Experiments on multiple datasets demonstrate the method's potential applicability.

Chapter 4 extends the concept of localized questions to Multimodal Large Language Models (MLLMs). The proposed targeted visual prompting involves creating a customized visual prompt containing the isolated region and the region in context. Visual components of the prompt are processed by a Swin Transformer and then projected into the input space of the LLM. Comprehensive experiments highlight the method's benefits across various medical VQA datasets.

Part II introduces two works in the field of consistency for VQA. The approach in Chapter 5 exploits the categorization of questions into perception and reasoning based on the visual abilities demanded from the model to answer them. This categorization informs a loss function term, enforcing consistency by penalizing inconsistent cases during training. The result is an improvement in both consistency and accuracy, showcasing the advantages of such model-agnostic approaches.

Chapter 6 builds upon the previous work by revising the definition of consistency and formalizing it from a more general perspective using logical implications. Similar to the prior method, a loss term is used to optimize consistency during training, encouraging the model to correct inconsistencies without compromising overall performance. Since implication annotations are usually not included in VQA datasets, we propose to predict them by leveraging a language model trained for the task of Natural Language Inference (NLI). Evaluation on medical and non-medical datasets supports the effectiveness of the approach compared to state-of-the-art consistency enhancement methods.

Finally, Part III contains a discussion and summary of the findings and limitations of the works presented in the thesis (Chapter 7), and offers possible directions for future work (Chapter 8).

2 Background

This chapter introduces key concepts related to Visual Question Answering (VQA). Given the multimodal nature of this task, we present concepts from NLP and computer vision separately, followed by a detailed exploration of their integration in VLMs. In the natural language section, we focus on RNNs and the transformer architecture, while for the computer vision section, we focus on CNNs and ViTs. Then, a historical approach is adopted to delve into VQA architectures and datasets.

2.1 Natural Language Processing, Understanding and Generation

This section delves into key concepts of AI applied to written language. It begins by clarifying the distinctions between NLP, Natural Language Understanding (NLU), and Natural Language Generation (NLG). Next, it examines essential processing steps like tokenization and word embeddings. Finally, the section explores crucial architectural developments like RNNs, Long Short-Term Memory (LSTM) networks, and transformers, concluding with a brief introduction to LLMs.

2.1.1 NLP vs. NLU vs. NLG

The topics of NLP, NLU and NLG, though related, are understood to mean different things. Broadly speaking, NLU and NLG are sub-topics of NLP (See Fig. 2.1), as described in the following definitions [42, 43]:

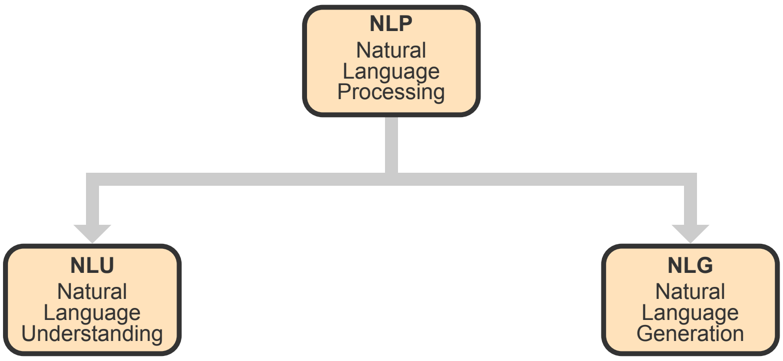

NLP
Natural Language Processing
NLU
Natural Language Understanding
NLG
Natural Language Generation

FIGURE 2.1: Relationship between NLP, NLU and NLG.

- **Natural Language Processing (NLP):** Rooted in computational linguistics, NLP comprises a wide range of operations applied to text. Its central aim is to add structure to text to endow computers with the ability to process it and generate responses.

- **Natural Language Understanding (NLU):** Delving deeper into textual meaning, NLU is concerned with the meaning of the text in terms of comprehension of grammar and context. Key features include part-of-speech tagging (adjectives, verbs, etc.), grammatical case recognition, and keyword identification.

- **Natural Language Generation (NLG):** Shifting the focus to generating text, NLG focuses on the generation of text in English or other languages. It encompasses tasks such as natural language generation, summarization, and translation.

2.1.2 Tokenization

Humans primarily use variable-length words arranged in sequences for language representation. Each word, encoded using standards like ASCII or UTF-8 [44], comprises a sequence of alphanumeric characters. Sentences and paragraphs remain unstructured data. To obtain structured data that is manipulable by computers, a process called tokenization breaks the text into discrete pieces. These pieces, known as tokens, can be words, subwords, or characters, depending on the desired granularity.

2.1.3 Word Embeddings

Word embeddings represent tokens as numerical vectors in a high-dimensional space. Given a sequence $\mathbf{T} = [w_1, w_2, ..., w_n]$ of n tokens, the embedding of the i-th word or token is denoted $E(w_i)$. The function E transforms the token w_i into a fixed-size, real-valued vector representation, allowing syntactically or semantically similar tokens to have comparable representations. Thus, the word embeddings can be represented as

$$\mathbf{WE} = [E(w_1), E(w_2), ..., E(w_n)] \in \mathbb{R}^{n \times J}, \tag{2.1}$$

where J is the dimension of each word embedding vector.

Some popular techniques to obtain word embeddings include Bag of Words (BoW) [45], Word2Vec [46], GloVe [47] and BERT [48].

2.1.4 Recurrent Neural Networks

RNNs [12] are networks that process sequential data. To understand the concept better, it is useful to start from the formulation of a dynamical system, as presented in [49], where a function f parameterized by $\boldsymbol{\theta}$ is applied to the previous state,

$$\mathbf{h}^{(t)} = f(\mathbf{h}^{(t-1)}; \boldsymbol{\theta}), \tag{2.2}$$

where $\mathbf{h}$ is the state of the system. This equation is said to be recurrent because the value of the state $\mathbf{h}$ at time t depends on its value at time $t-1$. For a given finite value of t, unfolding the graph that Eq. (2.2) represents is possible. This is, the equation is applied multiple times in a recurrent way to obtain a non-recurrent expression. For example, for $t = 3$,

$$\mathbf{h}^{(3)} = f(f(\boldsymbol{h}^{(1)}; \boldsymbol{\theta}); \boldsymbol{\theta}), \tag{2.3}$$

which reveals the function being applied multiple times in a sequential manner. This can be represented with a graph, as shown in Fig. 2.2, where each node represents a hidden state, and the edges represent the function.

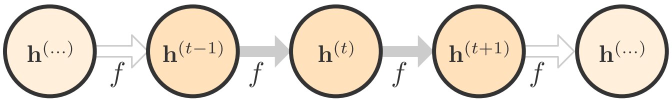


FIGURE 2.2: Unfolded graph for Eq. (2.2). Based on [49].

Including an external signal or input **s** results in a recurrent network, which can be represented by

$$\mathbf{h}^{(t)} = f(\mathbf{h}^{(t-1)}, \mathbf{s}^{(t)}; \boldsymbol{\theta}), \tag{2.4}$$

whose unfolded version is depicted in Fig. 2.3 with the outputs **o** for each state.

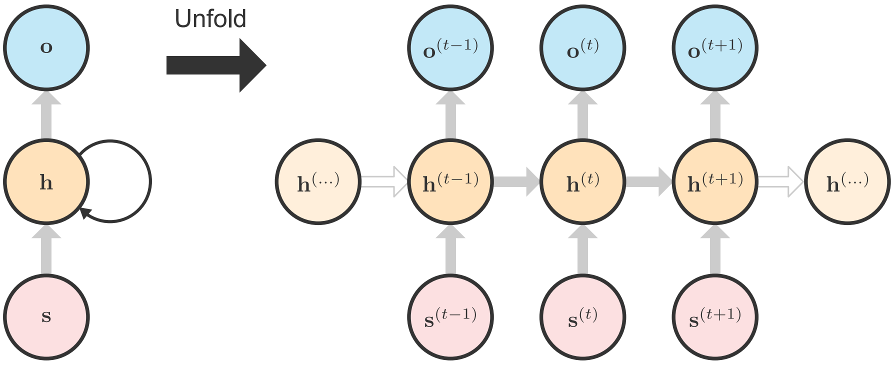


FIGURE 2.3: Unfolded RNN for Eq. (2.4). Based on [49].

One limitation of RNNs is the challenge in considering long-term dependencies in the input data. This issue arises from the exponentially smaller weights assigned to these long-term interactions (*i.e.*, vanishing gradient). For instance, consider the case in which the network is tasked with predicting the subsequent words in the sentence, "John is allergic to nuts. He refused to try the..." In this scenario, the context provided by the first sentence (nut allergy) aids in predicting more accurate words. However, if this context is situated at a greater distance from the word to be predicted, the network might struggle to utilize it effectively. A second issue is the occurrence of exploding gradients, leading to model instability and affecting the training process. Another constraint of the presented RNN is its unidirectional nature, limiting its capacity to incorporate future events for predicting more meaningful states [50]. We now present some variations that attempt to tackle these limitations.

RNN Variations

We examine the most common variations of RNNs

- **Bidirectional Recurrent Neural Network (BRNN)[51]:** In certain applications, like speech and handwriting recognition, generating an output that depends on all the input sequence elements can be beneficial. Bidirectional Recurrent Neural Networks (BRNNs) tackle this by merging two RNNs: one that begins at the initial sequence element and progresses forward, and another that commences at the final sequence element and progresses backward (refer to Fig. 2.4). This network configuration allows for outputs that take into account both past and future elements but are particularly sensitive to inputs near time step t.

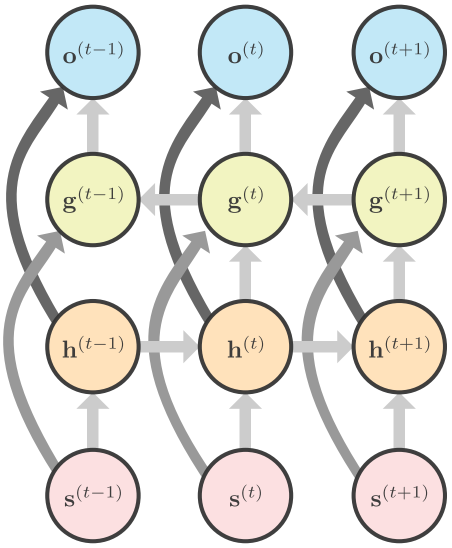


FIGURE 2.4: Unfolded BRNN. Based on [49].

- **Long Short-Term Memory (LSTM) [52]:** Arguably the most popular RNN architecture, it was introduced as a solution to the vanishing gradient issues of vanilla RNNs. Consequently, this architecture handles long-term dependencies more effectively. As illustrated in Fig. 2.5, the LSTM mitigates the long-term dependency problem by using a cell state and incorporating three types of gates: input, output, and forget. These gates facilitate control over the flow of information within the network. The cell, depicted in Fig. 2.5, comprises a cell state and a hidden state. The forget gate filters out irrelevant information from the previous cell state $\mathbf{C}^{(t-1)}$, such as a gender mentioned multiple times in the preceding sentences. Subsequently, the input gate determines which new information should be added to the current cell state $\mathbf{C}^{(t)}$. Finally, the output gate determines which information from the current cell state should be incorporated into the current hidden state $\mathbf{h}^{(t)}$.

  Equations (2.5)-(2.10) define the behaviour of each LSTM cell. Here, $\boldsymbol{U}_f$, $\boldsymbol{U}_i$, $\boldsymbol{U}_o$, $\boldsymbol{U}_g$, $\boldsymbol{W}_f$, $\boldsymbol{W}_i$, $\boldsymbol{W}_o$ and $\boldsymbol{W}_f$ correspond to learnable parameters of linear mappings, and $\odot$ is the element-wise product.

$$\boldsymbol{f}^{(t)} = \sigma\left(\mathbf{s}^{(t)}\boldsymbol{U}_f + \boldsymbol{h}^{(t-1)}\boldsymbol{W}_f\right) \tag{2.5}$$

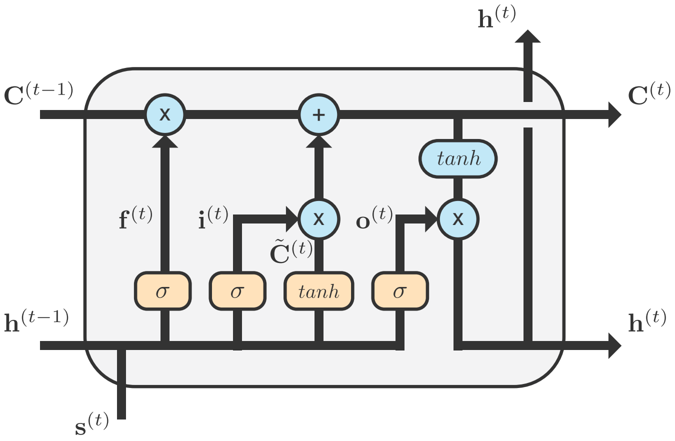


FIGURE 2.5: LSTM cell diagram.

$$\boldsymbol{i}^{(t)} = \sigma\left(\mathbf{s}^{(t)}\boldsymbol{U}_i + \boldsymbol{h}^{(t-1)}\boldsymbol{W}_i\right) \tag{2.6}$$

$$\boldsymbol{o}^{(t)} = \sigma\left(\mathbf{s}^{(t)}\boldsymbol{U}_o + \boldsymbol{h}^{(t-1)}\boldsymbol{W}_o\right) \tag{2.7}$$

$$\tilde{\boldsymbol{C}}^{(t)} = tanh\left(\mathbf{s}^{(t)}\boldsymbol{U}_g + \boldsymbol{h}^{(t-1)}\boldsymbol{W}_g\right) \tag{2.8}$$

$$\boldsymbol{C}^{(t)} = \sigma\left(\boldsymbol{f}^{(t)} \odot \boldsymbol{C}^{(t-1)} + \boldsymbol{i}^{(t)} \odot \tilde{\boldsymbol{C}}^{(t)}\right) \tag{2.9}$$

$$\boldsymbol{h}^{(t)} = tanh\left(\boldsymbol{C}^{(t)}\right) \odot \boldsymbol{o}^{(t)} \tag{2.10}$$

- **Gated Recurrent Unit (GRU) [53]:** This architecture also focuses on mitigating the long-term dependency issues of RNNs. Unlike the LSTM, Gated Recurrent Units (GRUs) do not make use of a cell state to control the information flow. It uses hidden states and has two gates (reset and update).

2.1.5 The Transformer Architecture

The transformer architecture [27] was introduced in 2017. Since its inception, this architecture has been applied in various domains, including language processing and computer vision. The utilization of transformers in image processing is further discussed in Sec. 2.2.2. As previously mentioned, recurrent neural networks have limitations, including the absence of parallelization, which hinders efficiency, and issues related to long-term dependencies due to vanishing and exploding gradients. The transformer addresses these limitations through its attention mechanism and architectural design.

The transformer, as depicted in Fig. 2.6, follows an encoder-decoder structure. In this setup, the input text undergoes mapping to a representation space by the encoder. Subsequently, the decoder utilizes this representation to sequentially generate an output sequence. This process is termed *auto-regressive* behavior because, at each time step, the previously generated elements serve as input for producing a new one.

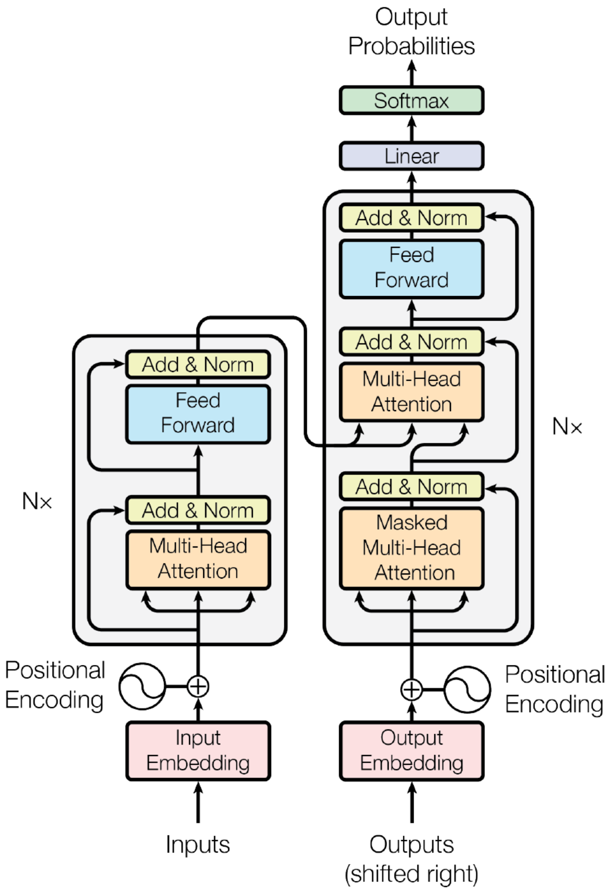


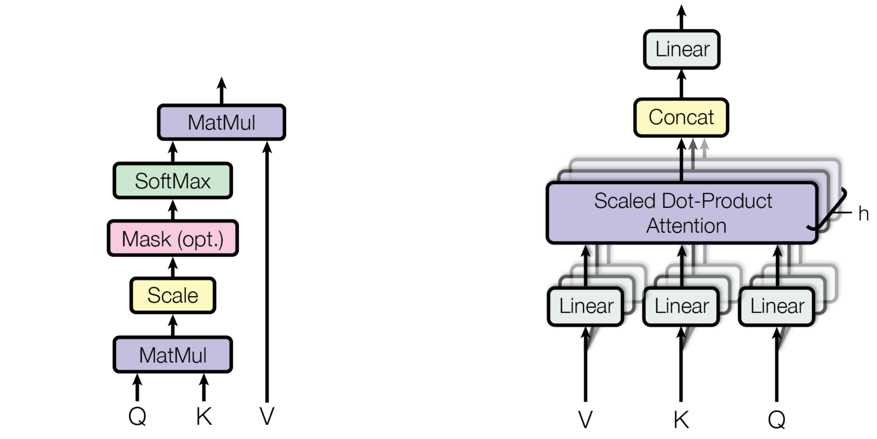
FIGURE 2.6: Transformer architecture. From [27].

Encoder

The encoder block, illustrated in Fig. 2.6, is responsible for generating a continuous representation from the embedded and positionally encoded inputs. The encoder consists of two sub-layers: a multi-head self-attention mechanism and a feed-forward network. Following [54], residual connections are incorporated into each sub-layer, along with layer normalization [55]. Instead of having one single encoder layer, the encoder is structured as a stack of six identical layers.

Decoder

The decoder block shares certain similarities with the encoder: It is constructed as a stack of six layers, employs residual connections and layer normalization, and incorporates both multi-head attention and a fully connected network. Nevertheless, it diverges by featuring two multi-head attention sub-layers instead of one. The second sub-layer serves the specific function of processing the output of the encoder. As depicted in Fig. 2.6, the sub-layer responsible for



(A) Scaled dot-product attention block.

(B) Multi-head attention.

FIGURE 2.7: Self-attention modules of the transformer architecture. From [27].

receiving the previously generated outputs incorporates a masking mechanism, preventing the model from attending to future positions.

Attention

The key part of the transformer architecture is its self-attention mechanism. The concept of *attention* was initially introduced in [56] for machine translation. The basic idea was to enable the model to determine which parts of the source sentence were relevant to predict the next word of the translation. In the transformer, the attention function is expressed in terms of three elements: queries, keys, and values, all represented as vectors. The output is computed by taking the weighted sum of the values, with each value assigned a weight determined by a compatibility function between the query and its corresponding key. Put differently, the transformer uses the self-attention mechanism to consider other words relevant to the word currently being processed. For instance, in translating the sentence "The cat climbed the bed because it was tired," self-attention allows the model to recognize that the word "it" is more closely associated with the word "cat" than with any other word in the sentence [57].

The transformer's implementation of the self-attention mechanism is called scaled dot-product attention and is illustrated in Fig. 2.7 (A). In this process, the dot products are computed between the query and all keys, then scaled by the dimension of the keys, d_k, and finally, a softmax function assigns weights of the values. This operation can be performed for a set of queries efficiently using matrices. Denoting the matrices $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ for queries, keys, and values, respectively, the attention operation is defined by Eq.(2.11).

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{2.11}$$

As shown in Fig. 2.7 (B), self-attention is performed for h “heads" $\{\boldsymbol{z_1}, \boldsymbol{z_2}, ..., \boldsymbol{z_h}\}$, projecting the queries, keys, and values with learned linear functions. This enables the simultaneous application of scaled dot-product attention on each head, producing output values with dimension d_v. Subsequently, these output values are concatenated and projected once again,

$$\text{MultiHead}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = Concat(\boldsymbol{z_1}, ..., \boldsymbol{z_h})\boldsymbol{W}^O \tag{2.12}$$

where

$$\boldsymbol{z_i} = \text{Attention}(\boldsymbol{Q}\boldsymbol{W_i^Q}, \boldsymbol{K}\boldsymbol{W_i^K}, \boldsymbol{V}\boldsymbol{W_i^V}) \tag{2.13}$$

with $\boldsymbol{W}$ representing the learnable parameters of the projection layers.

Positional Encoding

Positional encodings are necessary due to the lack of recurrence and convolutions so the model can consider the order of the input sequence. In the transformer architecture, positional encodings are added to the input embeddings for both the encoder and decoder blocks. Sine and cosine functions are used to this end, as follows

$$\boldsymbol{PE}_{(pos,2i)} = sin\left(\frac{\boldsymbol{pos}}{10000^{\frac{2i}{d_{\text{model}}}}}\right) \tag{2.14}$$

$$\boldsymbol{PE}_{(pos,2i+1)} = cos\left(\frac{\boldsymbol{pos}}{10000^{\frac{2i}{d_{\text{model}}}}}\right) \tag{2.15}$$

where $\boldsymbol{pos}$ represents the position and i the dimension. In other words, each dimension of the positional encoding represents a sinusoidal function, and the wavelengths exhibit a geometric progression ranging from 2π to $10000 \cdot 2\pi$.

2.1.6 Large Language Models

LLMs can be defined as very deep transformer-based models optimized for one or multiple language tasks using internet-scale datasets. These models can easily reach hundreds of billions of parameters. In fact, as illustrated in Fig. 2.8, there is a noticeable upward trend in model size. For instance, in June 2020, OpenAI unveiled GPT-3, which featured 175B parameters and was able to generate text and code with short prompts written by the user [58, 59]. One year later, Megatron-Turing NLG, with 530B parameters, was introduced [60]. The number of parameters for more recent models such as GPT-4 [61] has not been disclosed, but it is believed to exceed one trillion parameters.

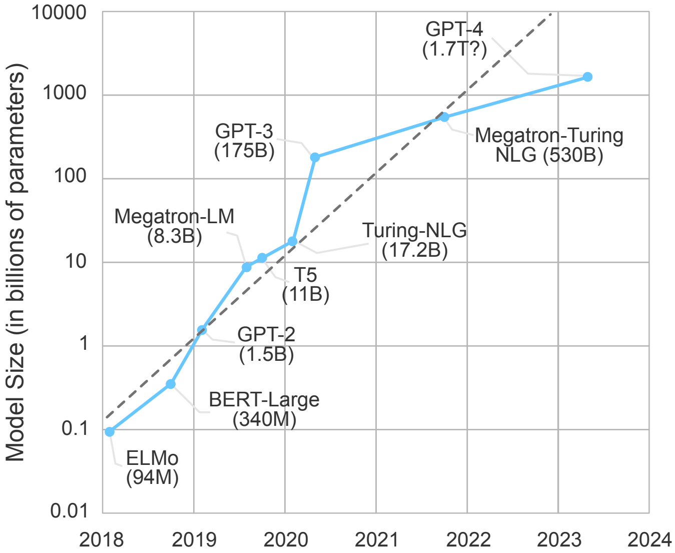


FIGURE 2.8: LLM size versus time. Adapted from [62].

Types

As mentioned in Sec. 2.1.5, the transformer architecture comprises an encoder and a decoder. Nevertheless, not all implementations based on this architecture adhere strictly to this design. In general, depending on the task at hand, LLM can be categorized into the following groups [59]:

- **Encoder only:** Models typically employed for tasks involving language understanding but without text generation as output. In such instances, the encoder is responsible for producing meaningful representations of the input, utilized by another network block, such as a classification head. Tasks like classification and sentiment analysis fall within this category. An example of this model type is BERT [48].

- **Decoder only:** These models are designed to generate high-quality language and content suitable for tasks such as blog generation and storytelling. GPT-3 [58] is an exemplar of this model type.

- **Encoder-decoder:** This model can both understand and generate text. Use cases include tasks like text translation and summarization. T5 [63] is an example of an architecture that utilizes an encoder-decoder structure.

Applications

LLMs can be used for diverse tasks. Fig. 2.9 shows the five main use case categories. For each category, some specific examples are provided.

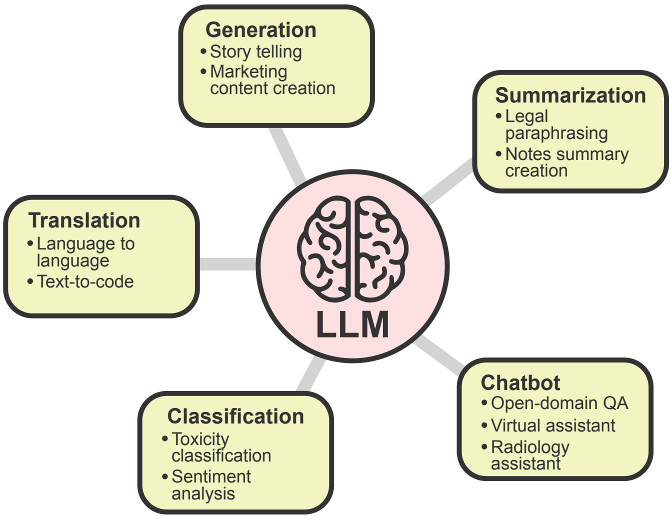


FIGURE 2.9: Applications of LLMs. Based on [59].

In the medical domain, LLMs have demonstrated diverse applications, including radiology report summarization [64], medical record evaluation [65], drug discovery [28], and others. Noteworthy breakthroughs have also been made in the field of robotics [66]. When equipped with vision capabilities (see Sec. 2.3.1), LLMs extend their applications to include radiology report generation [67], surgical training [68, 69], patient education [70], assistive technologies for visually impaired people [71], and autonomous driving [72], among others.

Challenges

Due to the large scale of both models and datasets, LLMs brings about special challenges that require consideration [59]:

- **Training cost:** The large scale of LLMs entails higher training requirements in terms of computing, capital and time, not only during development but also for deployment and maintenance. For example, the training of BLOOM, an open-source LLM with 176B parameter, took about 2.5 months, consumed 1,082,990 compute hours, and utilized 48 nodes with 8 Nvidia A100 80GB GPUs [73]. For GPT-3, it is estimated that the cost of training was over USD 12 million [74].
- **Scale of data:** LLMs require a substantial volume of data for training. In some cases, obtaining the data can be difficult due to privacy concerns (*e.g.*, medical data). In general, the curation of such extensive datasets poses a challenge in itself.
- **Technical expertise:** Given the large scale of the models, both training and deployment demand a certain level of expertise in areas such as deep learning pipelines, architectures, distributed computing, etc.

Due to these challenges, a set of methods has been proposed for fine-tuning LLMs, by minimizing the number of parameters that are modified. This provides advantages for researchers, institutions, and companies lacking the necessary infrastructure to properly fine-tune the model for a downstream task or a specific dataset. This set of methods is referred to as Parameter-Efficient Finetuning (PEFT). Some of the most popular techniques include Adapters [75], LoRA [76], and prefix tuning [77].

2.2 Computer Vision

This section briefly presents some essential computer vision concepts, with a focus on CNNs and ViTs, as they are the most relevant architectural types for this work.

2.2.1 Convolutional Neural Networks

CNNs are networks designed to process data represented in a 2D structure, such as images and time series. As its name indicates, CNNs employ a mathematical operation known as convolution, which is a specialized type of linear operation. Thus, CNNs can be characterized as neural networks that apply convolution in certain layers [49]. This operation takes place between the input and a convolution kernel. Rather than providing the continuous and 1D discrete definitions, we present the 2D convolution definition between an image **x** and a kernel **k**:

$$\boldsymbol{S}(i,j) = (\mathbf{x} * \boldsymbol{k})(i,j) = \sum_m \sum_n \mathbf{x}(m,n)\boldsymbol{k}(i-m, j-n) \tag{2.16}$$

where the indices m and n represent the valid positions for the operation. Eq. (2.16) indicates that the convolution operation is performed at every pixel location i, j of the image. With a kernel smaller than the input image, CNNs implement sparse interactions, implying that not every element of the output needs to be determined by every element of the input. This approach offers advantages in terms of efficiency and memory requirements.

CNNs are typically arranged in sequential convolutional layers, where the output of each layer undergoes processing by a subsequent layer. The output of each layer is commonly referred to as a feature map, as different layers extract image features at various levels of abstraction (such as edges, objects, etc.). Convolutional layers usually consist of three components:

- **Convolution block:** The input is convolved with h kernels to produce h feature maps.
- **Non-linearity block:** Applying a non-linearity or activation function enables the model to learn more complex functions. Some common non-linearities used in CNNs include Rectified Linear Unit (ReLU), leaky ReLU, and Gaussian Error Linear Unit (GELU) [78].
- **Pooling block:** Enables the reduction of the output size by summarizing each location

through a statistic involving several neighboring values. For instance, the max pooling operation [79] defines each value as the maximum value in a rectangular sub-region.

2.2.2 Vision Transformers

The Vision Transformer (ViT) extends the transformer architecture to handle images. As discussed in Sec. 2.1.5, the transformer operates on token vectorial representations, which are readily obtained for text through tokenization and token embeddings. The most effective equivalent of tokens for images and processing pipeline remained unclear until 2020 when the ViT was introduced [19].

The core idea of the ViT involves dividing the image into fixed-size tiles or patches, treating these as tokens, as depicted in Fig. 2.10. The patches need to be flattened first and then linearly projected to a constant latent vector dimension D. Then, positional embeddings are added to retain the information about the relative position of each patch. The result of this is processed by a transformer encoder block, which differs from the original transformer block in that the layer normalization is applied before every block instead of after (refer to Fig. 2.6). Since the ViT was introduced for image classification, the authors use a classification head at the model's output and leverage a learnable embedding prepended to the sequence of projected patches. This learnable embedding acts as image representation, following ideas proposed previously for BERT. Two popular versions of ViTs are Swin Transformers [80] and DeiT [81], which focus on scalability and data efficiency, respectively.

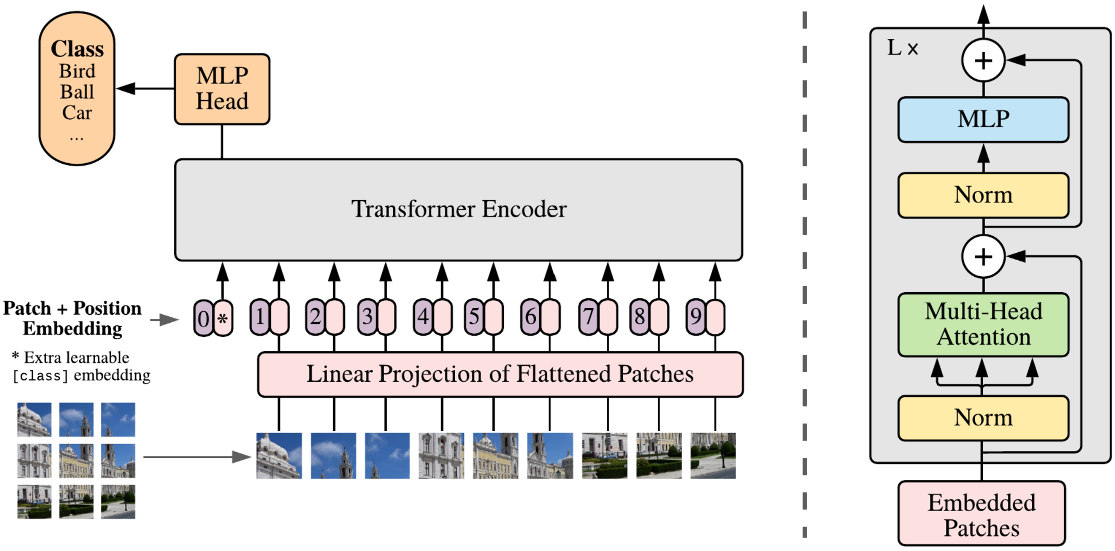


FIGURE 2.10: Vision Transformer (ViT) architecture overview. **Left:** General ViT pipeline. **Right:** Transformer encoder design. From [19].

2.3 Vision-Language Models

Models that integrate both vision and language capabilities are referred to as VLMs. The typical VLM structure consists of a vision encoder, a text encoder, and a strategy or mechanism to combine the two. During the training process using text and image data, the features of both modalities are expected to “align," signifying that the model learns representations that reveal the correspondence between text and vision, contributing to improved outputs. Tasks addressed by VLMs encompass:

- **Image retrieval:** Determining the most suitable image in a dataset that best corresponds to an input sentence [82].
- **Visual grounding:** Matching the words in an input sentence to the corresponding objects in an image [83].
- **Visual Question Answering:** Generating an appropriate answer to a question about an image [39]. Due to its significance in this work, a more detailed overview of this task is provided in Sec. 2.4.
- **Image captioning:** Given an image, providing an accurate caption that describes it [84].
- **Image-text labeling:** Assigning a label to an image-text tuple (*e.g.*, hate speech detection [85]).
- **Video summarization:** Generate a summary for an input video [86].

An important breakthrough in VLMs is CLIP [87], where vision and language models are trained end-to-end to maximize compatibility between matching image-caption tuples. This method is widely employed to pre-train vision encoders that can subsequently be utilized for other tasks.

2.3.1 Multimodal Large Language Models

A special sub-category of VLMs that has gained recent popularity is that of MLLMs [88–91], also referred to as MMLLM [92]. A simple definition of MLLMs is an LLM-based model capable of receiving and reasoning about one or more modalities different from text, with the ability to output text or any other modality. Under this definition, these models are not limited to images and language but can include other modalities such as audio, video, etc. Due to the relation with this work, we focus only on MLLMs in which the additional modality is an image and the output is text.

Two crucial questions arise at this point: (1) How to integrate the image into the LLM, and (2) How to train the model using multimodal inputs? Regarding the first question, two pathways have been explored: direct injection by aligning the image embeddings produced by a vision encoder to the LLM; and indirect injection, involving an expert proxy model translating the image into natural language [93]. Examples of the first case include models like BLIP-2 [94], LlaVA [89], InstructBLIP [88], VisionLLM [95], Otter [96], Llama-Adapters [97, 98]. An example of the second case is VideoChat [99]. In both situations, limitations in the vision model can propagate and affect the LLM output [100]. Regarding the second question, different techniques have been proposed to extend the understanding of LLMs to image data, of which the most relevant are the following [93]:

- **Visual instruction tuning (VIT) [89]:** An extension of instruction tuning [101], where a pre-trained LLM is fine-tuned on a set of instruction-formatted samples. This allows the model to learn to interpret instructions, execute them, and generalize to new instructions. The same principle is applied in visual instruction tuning, except that the image is also included in the instructions.

- **Visual In-Context Learning (ICL):** An extension of in-context learning [102], where the LLM is provided with a set of examples of the task at hand in different contexts, allowing generalization to new contexts. Visual ICL extends this concept by incorporating images into the examples.

- **Visual Chain-of-Thought:** An extension of Chain-of-thought, where the LLM is encouraged to reason through problems in a step-by-step manner. Visual Chain-of-Thought maintains the same behavior but incorporates images into the training prompts.

- **LLM-Aided Visual Reasoning (LAVR):** Inspired by works in which LLMs manipulate other models to execute tasks. In this scenario, the LLM can serve as a controller, a decision maker, or a semantics refiner [93].

2.4 Visual Question Answering

VQA is a multimodal task where a model provides an answer to a question about a given image [39]. Various skills are required from a VQA model to answer a question: finding relations, comparing objects, counting, perceiving visual features, etc. When the data has a medical nature, the task is referred to as Med-VQA. To differentiate, we refer to VQA for natural images as *general VQA*. This section presents some key aspects of VQA for both natural and medical images.

2.4.1 General VQA vs. Medical VQA

Although, in many cases, the principles, methods, and assumptions from VQA for natural images can be applied to Med-VQA, there are two specific challenges that make Med-VQA more complex:

- **Limited data:** As discussed later in Section 2.4.3, the size of Med-VQA datasets is considerably lower compared to general VQA. This is due to several reasons, such as the expense of data acquisition [103] and the need for specialized knowledge [104]. Med-VQA datasets require annotations from clinical experts who often lack sufficient time to generate the annotations that the task requires. Another reason is the privacy constraints that typically accompany medical data.

- **Uniqueness of medical images and vocabularies:** Medical data typically captures intricate information about the human body. Due to the wide variety of organs and tissues and the inter-patient variability and abnormalities, training models that accurately perform tasks on these images is challenging. It is often the case that images acquired by different machines have notable visual differences. All of this limits the development of, for instance, object detectors, which have been shown to benefit general VQA [105]. On the language side, the specialized vocabulary with relevant words that do not frequently appear in the text constitutes another obstacle for Med-VQA.

2.4.2 A Brief History of VQA

Fig. 2.11 provides a summary of the evolution of general and medical VQA architectures over time. The figure considers selected relevant publications, illustrating the overall progression of model structures and highlighting components that received more attention at different times.

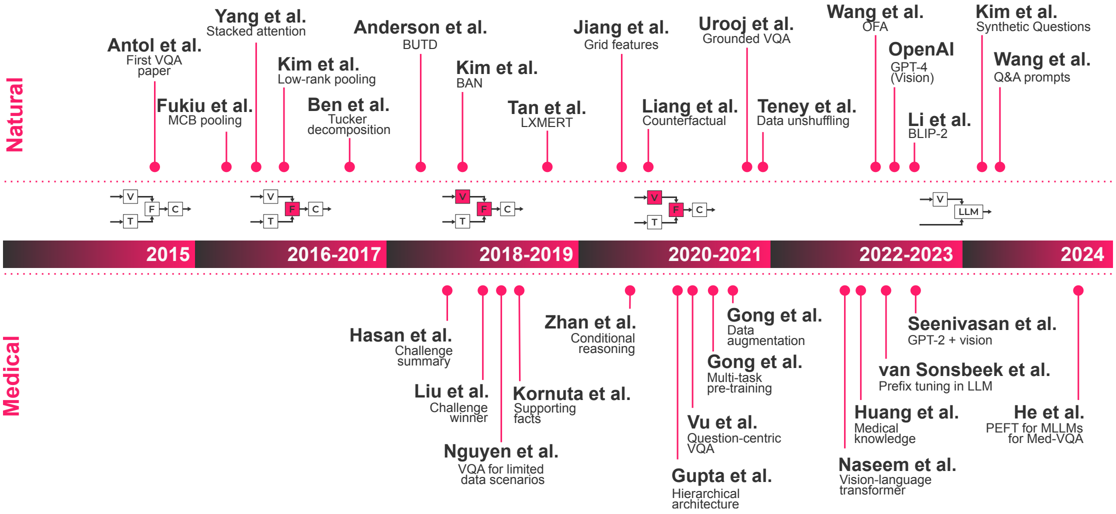


FIGURE 2.11: Evolution of VQA models for natural and medical images over time. Relevant publications are shown for each field. Above the year scale, schematic diagrams show the part(s) of the architecture that received the most focus at a given time. In the block diagrams, V stands for visual encoder, T for text encoder, F for multimodal fusion, and C for classifier.

General VQA

The VQA task was officially introduced as a challenge in 2015 by Antol et. al [39], building on previous works about visual queries [33, 36–38], as discussed in Sec. 1.1. The authors proposed the architecture shown in Fig. 2.12. This model follows the principle of generating embeddings separately for the image $\mathbf{x}$ and the question $\mathbf{q}$, projecting these to the same dimension, and then combining the projected embeddings using point-wise multiplication; the result of the product is then fed to a classifier, which selects the most likely answer $\hat{a}$, from a set of pre-defined answer $\mathcal{A}$. Mathematically, this can be formulated as

$$\hat{a} = \arg\max_{a \in \mathcal{A}} p(a|\mathbf{x}, \mathbf{q}; \boldsymbol{\theta}), \tag{2.17}$$

where $\boldsymbol{\theta}$ represents the parameters of the model, which is trained end-to-end using a cross-entropy loss.

With CNNs and LSTM networks being the standard for vision and text at the time, early research efforts in VQA focused on the multimodal fusion block. Pooling and decomposition techniques such as Multimodal Compact Bilinear (MCB) [106], Multimodal Low-rank Bilinear (MLB) [107], Multimodal Factorized Bilinear (MFB) [108], Multimodal Factorized High-order (MFH) [109], and Multimodal Tucker Fusion for Visual Question Answering (MUTAN) [110] were proposed. The idea behind these approaches is to facilitate richer interactions between the visual and text embeddings by using bilinear pooling [106–108] or Tucker decomposition [110], while simultaneously seeking low dimensionality to make the operations feasible at a large scale. Another breakthrough that happened at the time of the pooling methods was the concept of attention applied to VQA [111]. Here, through Stacked Attention Networks (SAN), the goal was to let the model learn which regions of the image were important to answer the question. This enabled spatially assigning different weights to the visual features to generate better answers and added some degree of explainability to the model.

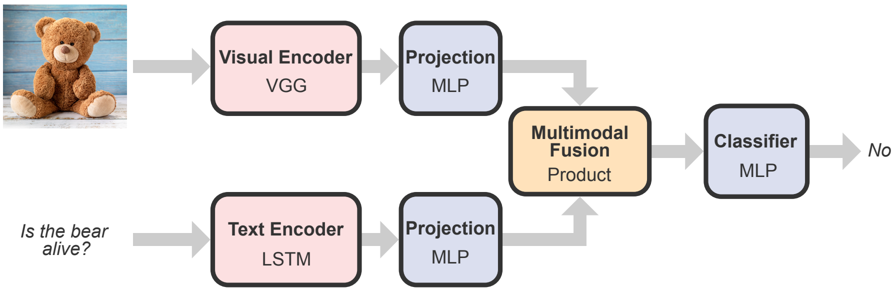


FIGURE 2.12: First VQA architecture. Image and question embeddings are extracted and then projected to the same dimension. The result is then combined using the element-wise product, and a classifier provides the most likely answer at the output.

Another breakthrough related to attention emerged with Bottom-Up Top-Down (BUTD) attention [112], where grid features produced by a CNN are replaced with object features produced by an Region-based Convolutional Neural Network (R-CNN) [113]. Attention is then computed along the region features to assign larger weights to the object regions that are more relevant to the question. Later on, Bilinear Attention Networks (BAN) builds on the object-based feature extraction but proposes adding co-attention [114] to MLB with the aim of considering the interaction between every object region and every question word.

One of the early signs hinting at the upcoming transition to transformer-based architectures for VQA was LXMERT [115]. This model comprises three encoder blocks: one for text, one for object relationships, and one to combine both modalities. The cross-modality block has a special design containing a bidirectional cross-attention sub-layer constructed from two uni-directional cross-attention sub-layers, one from text to vision and one from vision to text. LXMERT is notable for being trained on various tasks, including language modeling, masked object prediction, image-text matching, and VQA. Between 2020 and 2022, several approaches addressing different aspects of VQA were proposed. These include investigating the relevance of grid features against region features [116], learning with counterfactuals [117], data unshuffling [118], and visual grounding [119].

Subsequently, the OFA model [120] emerged as an encoder-decoder transformer with up to 930M parameters, serving, together with Flamingo [121], as precursors to MLLMs, such as GPT-4V [61], BLIP-2 [94], and Llama-adapters [97, 98]. A significant change introduced by these larger models is their ability to generate free-text answers instead of pre-defined ones, allowing for more varied responses and detailed descriptions.

As depicted in Fig. 2.11, MLLMs mark a paradigm shift in the fundamental structure of the VQA architecture. As discussed in Sec. 2.3.1, in numerous state-of-the-art approaches, visual features are extracted with a visual encoder, mapped to the dimension of the language tokens, and then integrated with the language tokens. This facilitates the seamless utilization of the LLM, which can also be left frozen during the training process [122]. More recent advancements in VQA focus on adding specialized world knowledge into the model [123] and using synthetic questions to answer human questions [124].

Medical VQA

In the medical domain, the evolution of Med-VQA has closely paralleled the progress made for natural images, with some approaches being directly adapted to the medical domain. Med-VQA is considered to have started later than general VQA (see Fig. 2.11), likely attributed, as mentioned earlier, to limited data availability and the associated annotation costs. The initiation of Med-VQA was significantly influenced by the ImageCLEF VQA-Med challenge, inaugurated in 2018, marking the initial applications of VQA to medical images. With a relatively small dataset with only 5,500 question-answer pairs for training, most of the approaches in the challenges were adapted from general VQA, such as SAN, MCB and MFB [125]. Addressing the challenge of limited data, a method was proposed in [126], utilizing meta-learning and a

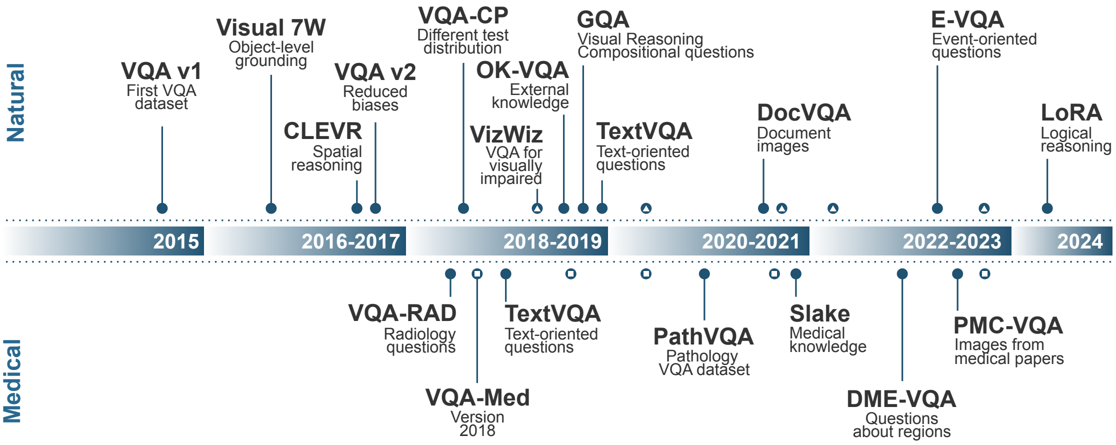


FIGURE 2.13: Evolution of VQA datasets for natural and medical images over time. Dots with triangles and squares indicate datasets produced for a specific VQA challenge.

Denoising Auto-Encoder (DAE) to generalize in limited-data scenarios and exploit unlabeled images, respectively.

Two methods were then introduced for generating answers using different modules based on the type of question asked, either open-ended or close-ended [105, 127]. These practices posed a significant challenge in the early years of Med-VQA, where certain methods were specifically tailored to the challenge data. Some approaches even treated the VQA problem as an image classification problem, disregarding the input questions [128]. Fortunately, as new datasets emerged, approaches shifted focus towards other aspects, including enhancing the importance assigned to questions [129] and incorporating data augmentation techniques [130]. From this point on, the adoption of transformer-based architectures, adapted from general VQA, became prominent in Med-VQA. This transition began with a pathology VQA model that employed a transformer to fuse text and visual features [131]. Later, the integration of MLLMs into Med-VQA has been observed [132–134].

2.4.3 Datasets

In terms of datasets, significant differences exist between general and medical VQA. Fig. 2.13 shows the most relevant datasets for natural and medical images over time. A notable characteristic of datasets in general VQA is the refinement of earlier versions, exemplified by the progression from VQA v1 [39] to VQA v2 [135]. Additionally, new versions like VQA-CP [136] were introduced to mitigate biases. This kind of dataset evolution is hardly observed in the medical domain, where, with some exceptions, there is only one version of the dataset. Due to the lack of data, challenges such as ImageCLEF VQA-Med have also re-used the same dataset used in previous versions [137].

Perhaps the most important dataset consideration is the number of images and question-answer pairs. As mentioned earlier, data collection in the medical domain is more challenging, resulting in substantially smaller datasets compared to their counterparts in general VQA.

To illustrate this, Table 2.1 provides an overview of publicly available VQA datasets for medical and natural images. Two conclusions can be drawn: (1) General VQA datasets tend to contain more images and more questions than Med-VQA datasets, and (2) Med-VQA datasets often incorporate automatically generated questions. Both of these considerations can be seen as consequences of the difficulties associated with medical data, as presented in Sec. 2.4.1. The first consideration impacts the quality of the models trained with such data, as achieving generalization becomes more challenging, making them more prone to biases. The second consideration limits the applicability or deployment of Med-VQA models in clinical environments, primarily due to the disparity between the nature of automatically generated questions and human-generated questions. Generally, automatically generated questions tend to adhere to a fixed structure that does not fully capture the semantic and syntactic variability and complexity of questions posed by humans.

2.5 Basics of Diabetic Macular Edema (DME) Staging

We briefly present basic concepts about fundus imaging and DME staging due to its relevance in this thesis. Fig. 2.14 (left) shows the basic anatomy of the eye. This organ exhibits a layered organization crucial for vision. Light initially passes through the transparent cornea, refracting and entering the anterior chamber filled with aqueous humor. The iris, the pigmented structure visible as eye color, regulates the incoming light via the central pupil. The crystalline lens, located behind the iris, further focuses the light onto the retina, the light-sensitive layer lining the posterior chamber. Within the retina, photoreceptor cells, known as rods and cones, convert light energy into electrical signals. These signals are then transmitted through the optic nerve to the visual cortex in the brain, allowing for visual perception. The entire globe of the eye is encased by the tough, white sclera, providing structural support and protection. This intricate interplay of structures enables the eye to capture and process visual information, transforming light into the world we see [138].

2.5.1 Fundus Imaging

In DME risk grade diagnosis from fundus images we are interested in capturing the rear of the eye, which is also known as fundus. This requires a specialized camera focused on the eye while emitting a bright light source, typically a flash or an infrared beam. The light travels through the eye, as described before, and reflects off the structures at the back of the eye and travels back through the pupil. A series of mirrors and lenses within the camera capture and concentrate this reflected light. Modern fundus cameras are digital, capturing an image directly onto a sensor instead of using film [157].

Dataset	# Images	# Questions	QA creation
E-VQA [139]	2,690	9,088	Automatic
OK-VQA [140]	14,031	14,055	Manual
VizWiz 2018 [34]	21,173	31,173	Manual
TextVQA [141]	28,408	45,336	Manual
DocVQA [142]	12,000	50,000	Manual
LoRA [143]	100,00	200,000	Automatic
Visual7W [144]	47,300	327,929	Manual
VQA-CPv1 [136]	205,000	370,000	Manual
VQAv1 [39]	204,721	614,163	Manual
VQA-CPv2 [136]	219,000	658,000	Manual
CLEVR [145]	100,000	864,968	Automatic
VQAv2 [135]	204,721	1’105,904	Manual
GQA [29]	113,000	22’000,000	Automatic
RadVisDial (gold) [146]	100	500	Manual
VQA-RAD [147]	316	3,515	Manual
VQA-Med 2020 [148]	5,000	5,000	Automatic
VQA-Med 2021 [137]	5,000	5,000	Automatic
VQA-Med 2018 [125]	2,866	6,413	Automatic
DME-VQA [149]	679	12,159	Automatic
Slake [150]	642	14,000	Manual
VQA-Med 2019 [151]	4,200	15,292	Automatic
VQA-Med 2023 [152]	5,000	25,000	Automatic
PathVQA [153]	4,998	32,799	Automatic
PMC-VQA [154]	149,000	227,000	Automatic
RadVisDial (silver) [146]	91,060	455,300	Automatic

TABLE 2.1: Overview of VQA dataset size sorted by the number of questions. **Top:** For natural images. **Bottom:** For medical images.

2.5.2 DME Staging

In assessing the severity of DME through color fundus images, a simplified classification system utilizes a three-point scale (0-2) to categorize disease progression. Grade 0 signifies a healthy retina, devoid of any visible "hard exudates," which appear as yellowish-white deposits. Grade 1 indicates the presence of these deposits confined to the peripheral regions of the retina, outside the central "macular area." Conversely, Grade 2 denotes the presence of hard exudates within the critical macular region, raising potential concerns for vision impairment. For practical purposes, the critical macular region is defined by a circle with a radius of one optic disc diameter [158]. Fig. 2.14 (right) shows an example where hard exudates are within this critical region, leading to Grade 2.

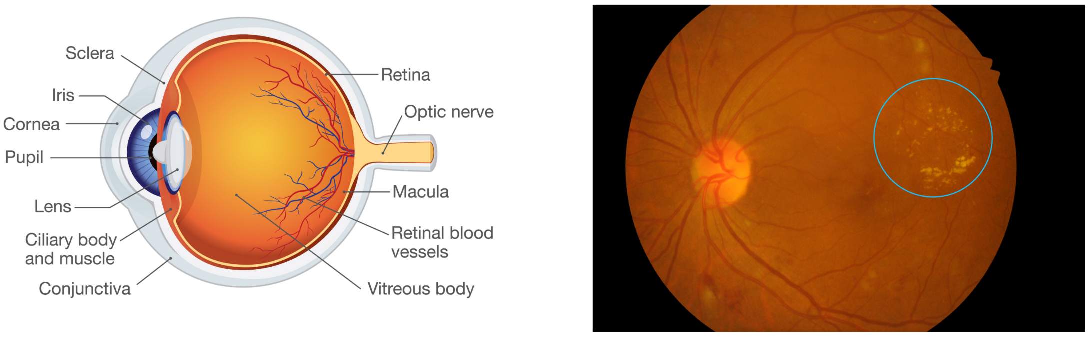


FIGURE 2.14: **Left:** Anatomy of the eye (from [155]). **Right:** Fundus image from the IDRiD dataset [156] with hard exudates encircled in light blue.

Enabling Localized Queries in VQA Part I

3 Localized Questions in Medical Visual Question Answering

The task of VQA has seen a relatively rapid development since it was first introduced back in 2015. With a few exceptions, VQA models have been applied to datasets with questions that refer to the entire image. This, however, can limit the interpretability of the model's predictions, as the model can benefit from biases in the data to produce the correct answer while disregarding the parts of the image that contain key information to answer the question. Furthermore, localized questions allow the comparison and quantification of agreement between questions about images and questions about regions. In this work, we present an attention-based method for medical VQA that enables the posing of questions about specific user-defined regions of an image while considering the context required to answer them. We benchmark our approach across multiple datasets and against different baselines, showing its effectiveness.

Author Contribution The contributing authors to this work are Pablo Márquez-Neila and Raphael Sznitman. My contributions to this chapter include the creation of the datasets, the development of the methodology, the conception and realization of the experiments, data analysis and interpretation, and visualization as well as the writing of the manuscript.

Publication This work is published in the Proceedings of the MICCAI 2023 conference [159].

License This work is published under the Springer License ↗. Reproduced with permission from Springer Nature.

3.1 Background and Previous Work

VQA models are neural networks that answer natural language questions about an image [29, 39, 115, 135]. The capability of VQA models to interpret natural language questions is of great appeal, as the range of possible questions that can be asked is vast and can differ from those used to train the models. This has led to many proposed VQA models for medical applications in recent years [103, 104, 125, 127, 129, 130, 160]. These models can enable clinicians to probe the model with nuanced questions, thus helping to build confidence in its predictions.

Recent work on Med-VQA has primarily focused on building more effective model architectures [129, 130, 161] or developing strategies to overcome limitations in Med-VQA datasets [129, 150, 162–164]. Another emerging trend is to enhance VQA performance by addressing the consistency of answers produced [149], particularly when considering entailment questions (*i.e.*, the answer to “Is the image that of a healthy subject?" should be consistent with the answer to “Is there a fracture in the tibia?"). Despite these recent advances, however, most VQA models are restricted to questions that consider the entire image at a time. Specifically, VQA typically uses questions that address content within an image without specifying where this content may or may not be in the image. Yet the ability to ask specific questions about regions or locations of the image would be highly beneficial to any user as it would allow fine-grained questions and model probing. For instance, Fig. 3.1 illustrates examples of such *localized questions* that combine content and spatial specifications. In the medical field, posing localized questions can significantly enhance the diagnostic process by providing second opinions to medical experts about suspicious regions. Additionally, this approach can improve trustworthiness by assessing the consistency between answers to both global and localized questions.

To this day, few works have addressed the ability to include location information in VQA models. In [165], localization information is posed in questions by constraining the spatial extent to a point within bounding boxes yielded by an object detector. The model then focuses its attention on objects close to this point. However, the method was developed for natural images and relies heavily on the object detector to limit the attention extent, making it difficult to scale in medical imaging applications. Alternatively, the approach from [129] answers questions about a pre-defined coarse grid of regions by directly including region information into the question (*e.g.*, “Is grasper in (0,0) to (32,32)?"). This method relies on the ability of the model to learn a spatial mapping of the image and limits the regions to be on a fixed grid. Localized questions were also considered in [149], but the region of interest was cropped before being presented to the model, assuming that the surrounding context is irrelevant for answering this type of question.

To overcome these limitations, we propose a novel VQA architecture that alleviates the mentioned issues. At its core, we hypothesize that by allowing the VQA model to access the entire images and properly encoding the region of interest, this model can be more effective at answering questions about regions. To achieve this, we propose using a multi-glimpse attention

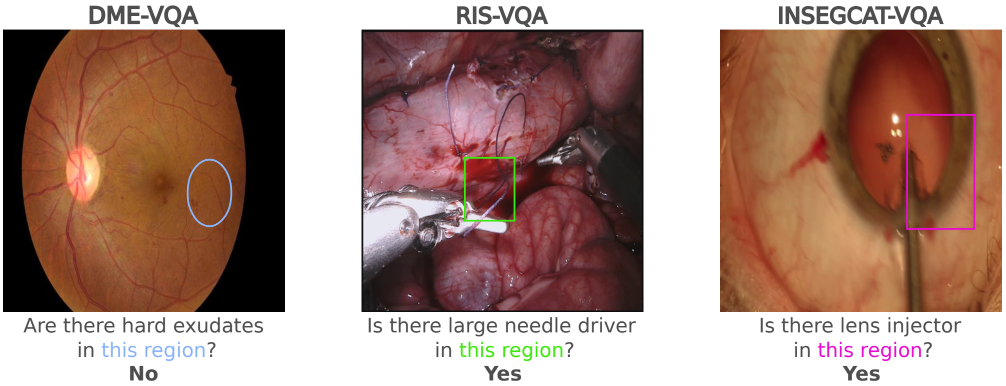


FIGURE 3.1: Examples of localized questions. In some cases (RIS-VQA and INSEGCAT-VQA), the object mentioned in the question is only partially present in the region. We hypothesize that context can play an important role in answering such questions.

mechanism [110, 129, 149] restricting its focus range to the region in question, but only after the model has considered the entire image. By doing so, we preserve contextual information about the question and its region. We evaluate the effectiveness of our approach by conducting extensive experiments on three datasets and comparing our method to state-of-the-art baselines. Our results demonstrate performance improvements across all datasets.

3.2 Method

Our method extends a VQA model to answer localized questions. We define a *localized question* for an image $\mathbf{x}$ as a tuple $(\mathbf{q}, \mathbf{m})$, where $\mathbf{q}$ is a question, and $\mathbf{m}$ is a binary mask of the same size as $\mathbf{x}$ that identifies the region to which the question pertains. Our VQA model, parameterized by $\boldsymbol{\theta}$ and depicted in Fig. 3.2, accepts an image and a localized question as input and produces a probability distribution over a finite set $\mathcal{A}$ of possible answers. The final answer $\hat{a}$ of the model is the element with the highest probability,

$$\hat{a} = \arg\max_{a \in \mathcal{A}} p(a \mid \mathbf{q}, \mathbf{x}, \mathbf{m}; \boldsymbol{\theta}). \tag{3.1}$$

The model proceeds in three stages to produce its prediction: input embedding, localized attention, and final classification.

3.2.1 Input Embedding

The question $\mathbf{q}$ is first processed by an LSTM [52] to produce an embedding $\hat{\mathbf{q}} \in \mathbb{R}^{Q}$. Similarly, the image $\mathbf{x}$ is processed by a ResNet-152 [54] to produce the feature map $\hat{\mathbf{x}} \in \mathbb{R}^{C \times H \times W}$.

3.2.2 Localized Attention

An attention mechanism uses the embedding to determine relevant parts of the image to answer the corresponding question. Unlike previous attention methods, we include the region information that the mask defines. Our *localized attention* module (Fig. 3.2 right) uses both descriptors and the mask to produce multiple weighted versions of the image feature map, $\hat{\mathbf{x}}' = \text{att}(\hat{\mathbf{q}}, \hat{\mathbf{x}}, \mathbf{m})$. To do so, the module first computes an attention map $\mathbf{g} \in \mathbb{R}^{G\times H\times W}$ with G glimpses by applying unmasked attention [129, 166] to the image feature map and the text descriptor. The value of the attention map at location (h, w) is computed as,

$$\mathbf{g}_{:hw} = \text{softmax}\left(\mathbf{W}^{(g)} \cdot \text{ReLU}\left(\mathbf{W}^{(x)}\hat{\mathbf{x}}_{:hw} \odot \mathbf{W}^{(q)}\hat{\mathbf{q}}\right)\right), \tag{3.2}$$

where the index $:hw$ indicates the feature vector at location (h, w), $\mathbf{W}^{(x)} \in \mathbb{R}^{C'\times C}$, $\mathbf{W}^{(q)} \in \mathbb{R}^{C'\times Q}$, and $\mathbf{W}^{(g)} \in \mathbb{R}^{G\times C'}$ are learnable parameters of linear transformations, and $\odot$ is the element-wise product. In practice, the transformations $\mathbf{W}^{(x)}$ and $\mathbf{W}^{(g)}$ are implemented with 1 × 1 convolutions and all linear transformations include a dropout layer applied to its input. The image feature maps $\hat{\mathbf{x}}$ are then weighted with the attention map and masked with $\mathbf{m}$ as,

$$\hat{\mathbf{x}}'_{cghw} = \mathbf{g}_{ghw} \cdot \hat{\mathbf{x}}_{chw} \cdot (\mathbf{m}\downarrow_{H\times W})_{hw}, \tag{3.3}$$

where c and g are the indexes over feature channels and glimpses, respectively, (h, w) is the index over the spatial dimensions, and $\mathbf{m}\downarrow_{H\times W}$ denotes a binary downsampled version of $\mathbf{m}$ with the spatial size of $\hat{\mathbf{x}}$. This design allows the localized attention module to compute the attention maps using the full information available in the image, thereby incorporating context into them before being masked to constrain the answer to the specified region.

3.2.3 Classification

The question descriptor $\hat{\mathbf{q}}$ and the weighted feature maps $\hat{\mathbf{x}}'$ from the localized attention are vectorized and concatenated into a single vector of size $C \cdot G + Q$ and then processed by a multi-layer perceptron classifier to produce the final probabilities.

3.2.4 Training

The training procedure minimizes the standard cross-entropy loss over the training set updating the parameters of the LSTM encoder, localized attention module, and the final classifier. The training set consists of triplets of images, localized questions, and the corresponding ground-truth answers. As in [39], the ResNet weights are fixed with pre-trained values, and the LSTM weights are updated during training.

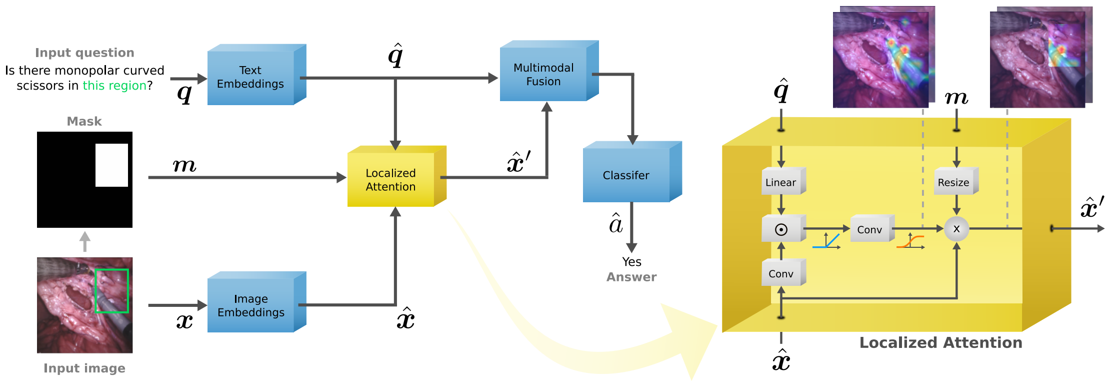


FIGURE 3.2: **Left:** Proposed VQA architecture for localized questions. The Localized Attention module allows the region information to be integrated into the VQA while considering the context necessary to answer the question. **Right:** Localized Attention module.

3.3 Experiments and Results

We compare our model to several baselines across three datasets and report quantitative and qualitative results. Additional results are available in Appendix A.

3.3.1 Datasets

We evaluate our method on three datasets containing questions about regions which we detail here. The first dataset consists of an existing retinal fundus VQA dataset with questions about the image's regions and the entire image. The second and third datasets are generated from public segmentation datasets but use the method described in [129] to generate a VQA version with region questions.

DME-VQA [149]: 679 fundus images containing questions about entire images (*e.g.*, "what is the DME risk grade?") and about randomly generated circular regions (*e.g.*, "are there hard exudates in this region?"). The dataset comprises 9'779 question-answer (QA) pairs for training, 2'380 for validation, and 1'311 for testing.

RIS-VQA: Images from the 2017 Robotic Instrument Segmentation dataset [167]. We automatically generated binary questions with the structure "is there [instrument] in this region?" and corresponding masks as rectangular regions with random locations and sizes. Based on the ground-truth label maps, the binary answers were labeled "yes" if the region contained at least one pixel of the corresponding instrument and "no" otherwise. The questions were balanced to maintain the same amount of "yes" and "no" answers. 15'580 QA pairs from 1'423 images were used for training, 3'930 from 355 images for validation, and 13'052 from 1'200 images for testing.

INSEGCAT-VQA: Frames of cataract surgery videos from the InSegCat 2 dataset [168]. We

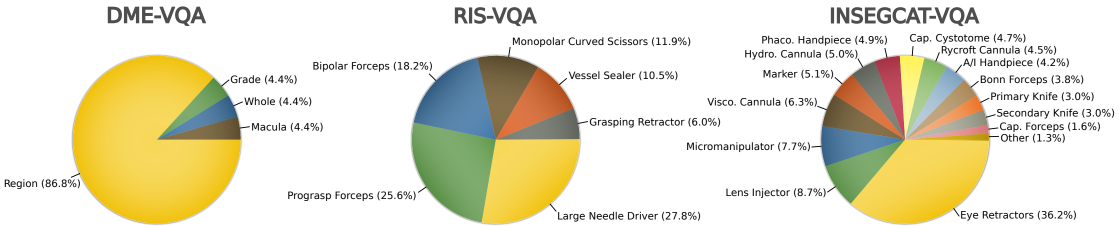


FIGURE 3.3: Distribution by question type (DME-VQA) and by question object (RIS-VQA and INSEGCAT-VQA).

followed the same procedure as in RIS-VQA to generate balanced binary questions with masks and answers. The dataset consists of 29'380 QA pairs from 3'519 images for training, 5'306 from 536 images for validation, and 4'322 from 592 images for testing.

Fig. 3.3 shows the distribution of questions in the three datasets.

3.3.2 Baselines and Metrics

We compare our method to four different baselines, as shown in Fig. 3.4:

No mask: no information is provided about the region in the question.

Region in text [129]: region information is included as text in the question.

Crop region [149]: image is masked to show only the queried region, with the area outside the region set to zero.

Draw region: region is indicated by drawing its boundary on the input image with a distinctive color.

We evaluated the performance of our method using accuracy for the DME-VQA dataset and the area under the Receiver Operating Characteristic (ROC) curve and Average Precision (AP) for the RIS-VQA and INSEGCAT-VQA datasets.

Implementation Details

Our VQA architecture uses an LSTM [52] with an output dimension 1024 to encode the question and a word embedding size of 300. We use the ResNet-152 [54] with ImageNet weights to encode images of size 448×448, generating feature maps with 2048 channels. In the localized attention block, the visual and textual features are projected into a 512-dimensional space before being combined by element-wise multiplication. Following [106, 166], the number of glimpses is set to $G = 2$ for all experiments. The classification block is a multi-layer perceptron

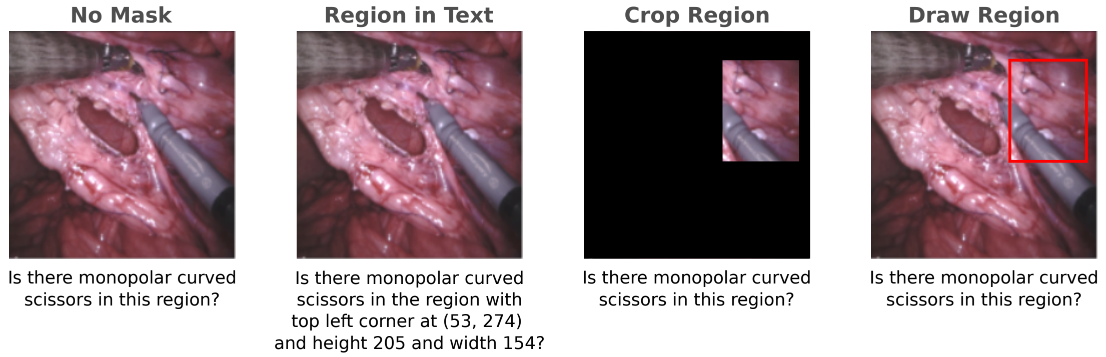

No Mask
Region in Text
Crop Region
Draw Region
Is there monopolar curved scissors in this region?
Is there monopolar curved scissors in the region with top left corner at (53, 274) and height 205 and width 154?
Is there monopolar curved scissors in this region?
Is there monopolar curved scissors in this region?

FIGURE 3.4: Illustration of the evaluated baselines for an example image.

Method	Accuracy (%)				
	Overall	Grade	Whole	Macula	Region
No Mask	61.1 ± 0.4	80.0 ± 3.7	85.7 ± 1.2	**84.3** ± **0.5**	57.6 ± 0.4
Region in Text [129]	60.0 ± 1.5	57.9 ± 12.5	85.1 ± 1.9	83.2 ± 2.4	57.7 ± 1.0
Crop Region [149]	81.4 ± 0.3	78.7 ± 1.3	81.3 ± 1.7	82.3 ± 1.4	81.5 ± 0.3
Draw Region	83.0 ± 1.0	79.6 ± 2.5	77.0 ± 4.8	84.0 ± 1.9	83.5 ± 1.0
Ours	**84.2** ± **0.6**	**82.8** ± **0.4**	**87.0** ± **1.2**	83.0 ± 1.5	**84.2** ± **0.7**

TABLE 3.1: Average accuracy for different methods on the DME-VQA dataset. The results shown are the average of 5 models trained with different seeds.

with a hidden layer of 1024 dimensions. A dropout rate of 0.25 and ReLU activation are used in the localized attention and classifier blocks.

We train our models for 100 epochs using an early stopping condition with patience of 20 epochs. Data augmentation consists of horizontal flips. We use a batch size of 64 samples and the Adam optimizer with a learning rate of 10^{-4}, which is reduced by a factor of 0.1 when learning stagnates. Models implemented in PyTorch 1.13.1 and trained on an Nvidia RTX 3090 graphics card[1].

3.3.3 Results

Our method outperformed all considered baselines on the DME-VQA (Table 3.1), the RIS-VQA, and the INSEGCAT-VQA datasets (Table 3.2), highlighting the importance of contextual information in answering localized questions. Context proved to be particularly critical in distinguishing between objects of similar appearance, such as the bipolar and prograsp forceps in RIS-VQA, where our method led to an 8 percent point performance improvement (Table 3.3). In contrast, the importance of context was reduced when dealing with visually

[1] Our code and data are available at https://github.com/sergiotasconmorales/locvqa

Dataset	Method	AUC	AP
RIS-VQA	No Mask	0.500 ± 0.000	0.500 ± 0.000
	Region in Text [129]	0.677 ± 0.002	0.655 ± 0.003
	Crop Region [149]	0.842 ± 0.002	0.831 ± 0.002
	Draw Region	0.835 ± 0.003	0.829 ± 0.003
	Ours	**0.885** ± **0.003**	**0.885** ± **0.003**
INSEGCAT-VQA	No Mask	0.500 ± 0.000	0.500 ± 0.000
	Region in Text [129]	0.801 ± 0.012	0.793 ± 0.014
	Crop Region [149]	0.901 ± 0.002	0.891 ± 0.003
	Draw Region	0.910 ± 0.003	0.907 ± 0.005
	Ours	**0.914** ± **0.002**	**0.915** ± **0.002**

TABLE 3.2: Average test AUC and AP for different methods on the RIS-VQA and INSEGCAT-VQA datasets. The results shown are the average over 5 seeds.

Method	Instrument Type					
	Large Needle Driver	Monopolar Curved Scissors	Vessel Sealer	Grasping Retractor	Prograsp For-ceps	Bipolar Forceps
No Mask	0.500 ±0	0.500 ±0	0.500 ±0	0.500 ±0	0.500 ±0	0.500 ±0
Region in Text [129]	0.717 ±0.003	0.674 ±0.001	0.620 ±0.011	0.616 ±0.020	0.647 ±0.008	0.645 ±0.003
Crop Region [149]	0.913 ±0.002	0.812 ±0.003	0.752 ±0.009	0.715 ±0.015	0.773 ±0.003	0.798 ±0.004
Draw Region	0.915 ±0.003	0.777 ±0.003	0.783 ±0.004	0.709 ±0.012	0.755 ±0.004	0.805 ±0.005
Ours	**0.944 ±0.001**	**0.837 ±0.005**	**0.872 ±0.008**	**0.720 ±0.031**	**0.834 ±0.006**	**0.880 ±0.003**

TABLE 3.3: Average test AUC for different methods on the RIS-VQA dataset as a function of instrument type. Results are averaged over 5 models trained with different seeds. The corresponding table for INSEGCAT-VQA is available in Appendix A.

distinct objects, resulting in smaller performance gains as observed in the INSEGCAT-VQA dataset. For example, despite not incorporating contextual information, the baseline *crop region* still benefited from correlations between the location of the region and the instrument mentioned in the question (*e.g.*, the eye retractor typically appears at the top or the bottom of the image), enabling it to achieve competitive performance levels that are less than 2 percent points lower than our method (Table 3.2, bottom).

Similar to our method, the baseline *draw region* incorporates contextual information when

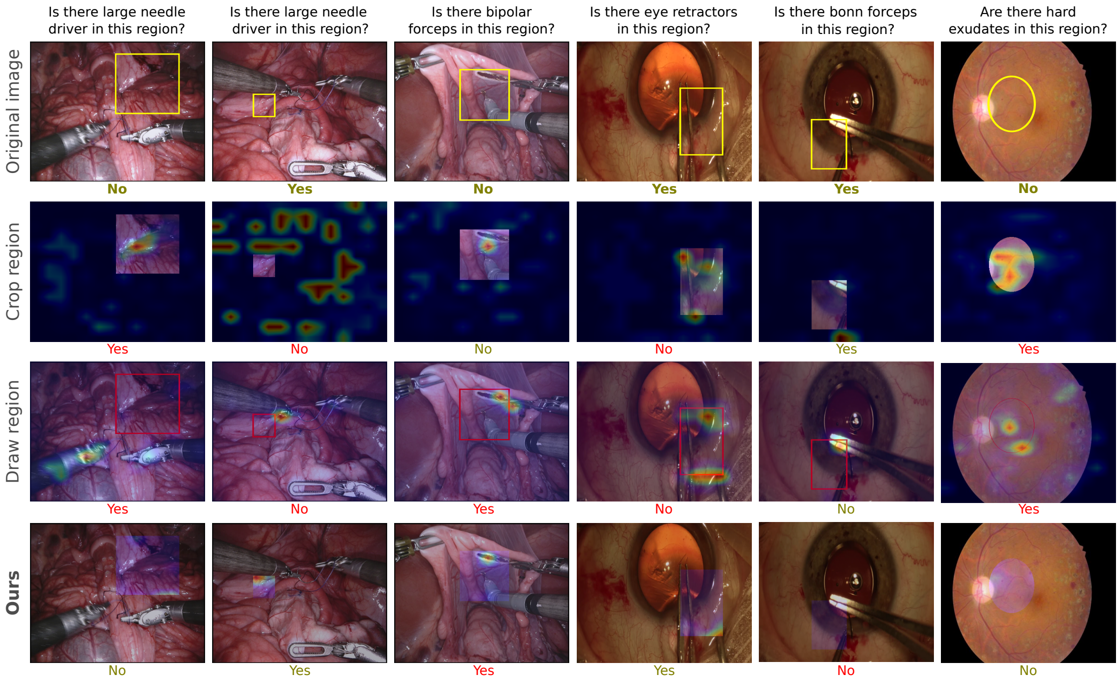

Is there large needle driver in this region?
Is there large needle driver in this region?
Is there bipolar forceps in this region?
Is there eye retractors in this region?
Is there bonn forceps in this region?
Are there hard exudates in this region?
Original image
No
Yes
No
Yes
Yes
No
Crop region
Yes
No
No
No
Yes
Yes
Draw region
Yes
No
Yes
No
No
Yes
Ours
No
Yes
Yes
Yes
No
No

FIGURE 3.5: Qualitative examples on the RIS-VQA dataset (columns 1-3), INSEGCAT-VQA (columns 4-5), and DME-VQA (last column). Only the strongest baselines were considered in this comparison. The first row shows the image, the region, and the ground truth answer. Other rows show the overlaid attention maps and the answers produced by each model. Wrong answers are shown in red.

answering localized questions. However, we observed that drawing regions on the image can interfere with the computation of guided attention maps, leading to incorrect predictions (Fig. 3.5, column 4). In addition, the lack of masking of the attention maps often led the model to wrongly consider areas beyond the region of interest while answering questions (Fig. 3.5, column 1).

When analyzing mistakes made by our model, we observe that they tend to occur when objects or background structures in the image look similar to the object mentioned in the question (Fig. 3.5, column 3). Similarly, false predictions were observed when only a few pixels of the object mentioned in the question were present in the region.

3.4 Conclusion

In this work, we proposed a novel VQA architecture to answer questions about regions. We compare the performance of our approach against several baselines and across three different datasets. By focusing the model's attention on the region after considering the evidence in the full image, we show how our method brings improvements, especially when the complete image context is required to answer the questions. Future works include studying the agreement between answers to questions about concentric regions, as well as the agreement between questions about images and regions.

4 Targeted Visual Prompting for Medical Visual Question Answering

With growing interest in recent years, Med-VQA has rapidly evolved with MLLMs emerging as an alternative to classical model architectures. Specifically, their ability to add visual information to the input of pre-trained LLMs brings new capabilities for image interpretation. However, simple visual errors cast doubt on the actual visual understanding abilities of these models. To address this, region-based questions have been proposed as a means to assess and enhance actual visual understanding through compositional evaluation. To combine these two perspectives, this work introduces targeted visual prompting to equip MLLMs with region-based questioning capabilities. By presenting the model with both the isolated region and the region in its context in a customized visual prompt, we show the effectiveness of our method across multiple datasets while comparing it to several baseline models.

Author Contribution Co-authored alongside Raphael Sznitman and Plablo Márquez-Neila, my contributions to this project involved creating datasets, formulating methodologies, designing experiments, analyzing and visualizing results, and composing the manuscript.

Publication This work has been accepted to the AMAI Workshop of the MICCAI 2024 conference.

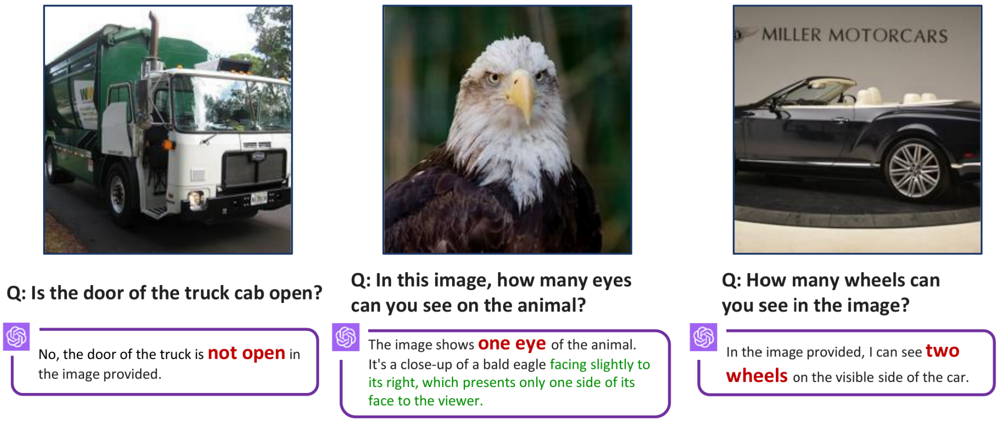
License Accepted and soon to be published; Applications of Medical Artificial Intelligence, 2024, by Springer Nature under the Springer License ↗. Reproduced with permission from Springer Nature.

4.1 Background and Previous Work

VQA is centered on developing models capable of answering questions about specific images [39]. This task is particularly challenging within the medical domain due to factors such as a scarcity of annotated data [103, 162], the wide variety of imaging modalities and anatomical regions [105], as well as the unique characteristics of medical images and terminology, all of which necessitate specialized expertise [103, 127]. Furthermore, approaches that leverage the detection of natural objects, which have significantly improved performance in the analysis of natural images [112], are less straightforward when applied to medical imagery [105].

Historically, models for Med-VQA treated visual and textual information independently, later merging these features through various fusion techniques. This composite data would then be input into a classifier to determine the most probable answer. However, recent developments in transformer-based models [27], including advancements in LLMs, have led to a notable shift in VQA strategies. These advancements have paved the way for the adoption of MLLMs that integrate both visual and textual data more seamlessly, a trend that is emerging in both general [93, 100, 169] and medical VQA [133, 154] applications.

Despite the remarkable adoption of MLLMs, recent research has raised concerns about the quality of their visual capabilities (Fig. 4.1). This issue primarily arises from the pre-training process of the visual component, which typically relies on models like CLIP [87]. Surprisingly, MLLMs can perceive certain visually distinct images as similar, a phenomenon that human observers readily recognize as different [100]. These visual understanding failures were also observed in VQA models before the widespread adoption of MLLMs [29, 40, 135, 170].



Q: Is the door of the truck cab open?

No, the door of the truck is **not open** in the image provided.

Q: In this image, how many eyes can you see on the animal?

The image shows **one eye** of the animal. It's a close-up of a bald eagle facing slightly to its right, which presents only one side of its face to the viewer.

Q: How many wheels can you see in the image?

In the image provided, I can see **two wheels** on the visible side of the car.

FIGURE 4.1: Examples of visual understanding failures using GPT-4V for the VQA task. From [100].

To enhance explainability in the visual component of Med-VQA, the work in [159] proposes a novel approach using the formulation of *localized questions* [159]. These questions allow fine-grained probing of images by focusing on user-defined regions rather than the entire image and facilitate a *compositional evaluation* of the model's reasoning abilities. To enable such localized questions, the region to query is encoded and directly integrated into the attention mechanism of the model. By doing so, the model gains access to context relevant to answering the question. Alternatively, other proposed strategies include providing the model with a restricted region of the image [149] or relying on the language component of the VQA model to interpret region coordinates directly included in the question [129]. Notably, [165] limits spatial focus by considering only certain bounding boxes produced by an object detector. Yet all of these methods suffer from the same common limitation: MLLMs cannot directly be integrated into these schemes to leverage their benefits for Med-VQA.

To this end, we propose to overcome this challenge by introducing a novel approach, namely *Targeted Visual Prompting*. By carefully designing a prompt that provides both global and local visual tokens relative to the region of interest defined by the user, our method allows the full advantage of the MLLM to enhance the performance of the VQA model. To validate the effectiveness of our method, we conduct exhaustive experiments across multiple datasets. Our results demonstrate clear performance benefits compared to previously proposed methods, all achieved without introducing additional parameters to the model.

4.2 Method

In general, a VQA model with parameters $\boldsymbol{\theta}$ generates an answer $\hat{a}$ when given an input image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ and a related question represented as a sequence of words, $\mathbf{q}$. In its most general form, this process can be described as a function Ψ_{VQA}, parameterized by $\boldsymbol{\theta}$, that is applied on the image-question pair,

$$\hat{a} = \Psi_{\text{VQA}}(\mathbf{x}, \mathbf{q}; \boldsymbol{\theta}). \tag{4.1}$$

Traditionally, this model's output is a distribution over a set of N candidate answers $\{a_1, a_2, ..., a_N\}$ set beforehand.

In this work, however, we choose the answer of the VQA to be generated by an LLM in an auto-regressive manner until the end-of-sentence (EOS) token is produced. To make the LLM multimodal, we adopt the widely used approach of projecting visual embeddings onto the input space of the LLM [67, 89, 171] and express this as

$$\hat{a} = \Psi_{\text{LLM}}(\Psi_{\text{Vis}}(\mathbf{x}, \boldsymbol{\theta}_{\text{Vis}})\mathbf{W}^{\text{proj}}, \mathbf{q}; \boldsymbol{\theta}_{\text{LLM}}), \tag{4.2}$$

where Ψ_{Vis} refers to the visual encoder with parameters $\boldsymbol{\theta}_{\text{Vis}}$, and $\mathbf{W}^{\text{proj}}$ denotes the learnable parameters of the projection layer. Although not explicitly formalized, it is implied that the answer is generated in an auto-regressive fashion, meaning that the next word in the answer depends on the previously predicted words.

To expand the model's capability to handle localized questions, we propose here a dedicated targeted visual prompt that allows two perspectives of the image to be encoded: one containing only the region of the image and the other containing the region in context.

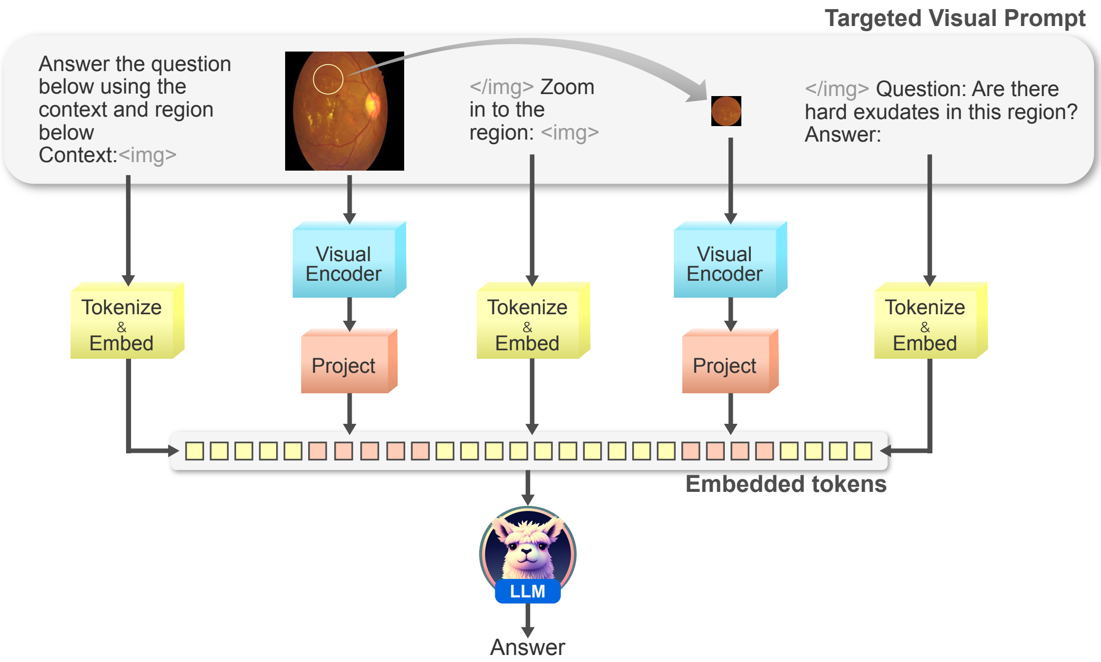

Targeted Visual Prompt
Answer the question below using the context and region below Context:<img>
</img> Zoom in to the region: <img>
</img> Question: Are there hard exudates in this region? Answer:
Visual Encoder
Project
Tokenize & Embed
Visual Encoder
Project
Tokenize & Embed
Tokenize & Embed
Embedded tokens
LLM
Answer

FIGURE 4.2: Our customized targeted visual prompt is created by providing the model with the region in context, as well as an isolated version of the region. Visual tokens are projected to the input space of the LLM and concatenated with the instruction tokens.

The targeted visual prompt consists of five components: (1) comprises model instruction, denoted as $\mathbf{w}_{\text{instr}}$; (2) the visual context represented by the image with the region drawn on it, $\mathbf{x}_r$; (3) $\mathbf{w}_{\text{det}}$ contains a textual prefix for the region; (4) the cropped region $\mathbf{r}$; and (5) $\mathbf{w}_q$ includes the question $\mathbf{q}$. Text-containing parts of the prompt undergo tokenization and embedding, while the visual components are processed by a visual encoder and then projected into the input space of the LLM. Subsequently, the results are concatenated and processed by the LLM, resulting in the generation of an answer. To handle global questions, the entire image is assigned to $\mathbf{r}$. We illustrate our model in Fig. 4.2 and summarize the computation of the answer as,

$$\hat{a} = \Psi_{\text{LLM}}(\mathbf{w}_{\text{instr}}, \Psi_{\text{Vis}}(\mathbf{x}_r, \boldsymbol{\theta}_{\text{Vis}})\mathbf{W}^{\text{proj}}, \mathbf{w}_{\text{det}}, \Psi_{\text{Vis}}(\mathbf{r}, \boldsymbol{\theta}_{\text{Vis}})\mathbf{W}^{\text{proj}}, \mathbf{w}_q; \boldsymbol{\theta}_{\text{LLM}}), \tag{4.3}$$

Training. As in [67], our model is trained using the original auto-regressive training loss of the LLM. The loss function is the standard negative log-likelihood accumulated over all time steps for predicting the correct next token. For a ground truth answer of length T, this loss is expressed as,

$$\mathcal{L}(\boldsymbol{\theta}) = -\sum_{t=1}^{T} \log p_\theta(a^t | \mathbf{x}, \mathbf{w}, a^{1:t-1}; \boldsymbol{\theta}), \tag{4.4}$$

where $\mathbf{x}$ and $\mathbf{w}$ denote the visual and textual elements, respectively, and $\mathbf{a} = \{a_1, a_2, ..., a_T\}$ is the ground truth answer.

4.3 Experiments and Results

4.3.1 Datasets

To evaluate our method, we make use of several publically available datasets: (1) DME-VQA: contains questions on DME risk grade and about the presence of biomarkers in the entire image or specific regions. (2) RIS-VQA: contains images from the DaVinci robot during gastrointestinal surgery and questions related to surgical instruments. (3) INSEGCAT-VQA: contains frames from cataract surgery videos and questions about instruments used in this type of surgery. A summary of these is shown in Table 4.1, based on [159].

Dataset	Modality	# images	# QA-pairs
DME-VQA	Fundus	679	13470
RIS-VQA	Gastrointestinal	2978	32562
INSEGCAT-VQA	Cataract surgery	4647	39008

TABLE 4.1: Main parameters of the used datasets.

4.3.2 Baselines

We benchmark our method against multiple baselines, which are exemplified in Fig. 4.3. In **No mask**, the model receives no information about the location of the region; in **Region in text**, the region is specified in the question; in **Draw region**, the region is marked on top of the image. In **Context only**, the model only sees the context, but not the contents of the region; in **Crop region**, the model receives no context; finally, in **LocAtt** [159], the model has access to the image, as well as a binary image representing the region. For these baselines, the visual prompt given to the model is: *"Answer the question below using the context below Context: <Img><Image></Img> Question:<Question> Answer:"*

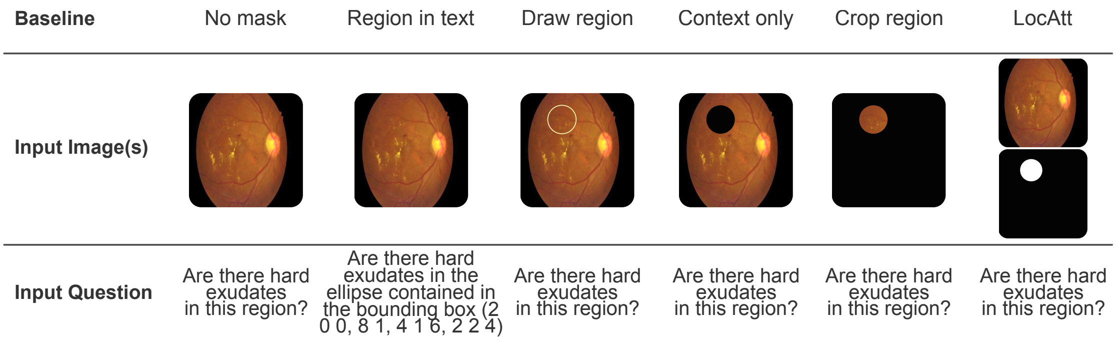

Baseline
No mask
Region in text
Draw region
Context only
Crop region
LocAtt
Input Image(s)
Input Question
Are there hard exudates in this region?
Are there hard exudates in the ellipse contained in the bounding box (2 0 0, 8 1, 4 1 6, 2 2 4)
Are there hard exudates in this region?
Are there hard exudates in this region?
Are there hard exudates in this region?
Are there hard exudates in this region?

FIGURE 4.3: Example input images and questions for evaluated baselines. In the baseline "Region in text," the digits are separated to provide a fair scenario to the LLM.

4.3.3 Implementation Details

We use R2GenGPT [67] as base MLLM, adapting it from the task of radiology report generation to VQA. Following the original implementation of R2GenGPT, we use a pre-trained Swin Transformer [80] as a visual encoder and Llama 2 [172] as LLM initialized with its official weights. Different from to R2GenGPT, we finetune all modules, including the LLM, end-to-end. We use the default parameters for the selected backbone model: We train all our models for 15 epochs, with a batch size of 8 and a learning rate of 1e-4, with the AdamW optimizer and a cosine annealing scheduler with a minimum learning rate of 1e-6. For the text generation, we use a repetition penalty of 2.0 as in [67] but establish a length penalty of -1.0 to encourage short answers. Our implementation uses PyTorch 2.0.1 and two Nvidia A100 cards with 80 GB of memory each.

4.3.4 Results

Table 4.2 summarizes our results on the DME-VQA, RIS-VQA, and INSEGCAT-VQA datasets. The accuracy and F1 score are reported for all datasets. Notably, our method consistently outperforms all evaluated baselines across all datasets, underscoring the efficacy of targeted visual prompting in enhancing MLLMs with region-based capabilities.

In the case of the DME-VQA and RIS-VQA datasets, we observe that the performance of *context only* surpasses that of *crop region*. At first glance, this suggests that the context holds more relevance than the specific contents of the region. However, this behavior is likely influenced by spurious correlations between region sizes, locations, and the corresponding answers. For instance, in DME-VQA, images with a high amount of biomarkers often feature smaller regions associated with negative answers. Similarly, in RIS-VQA, the tool can often be determined from its body without considering the tip.

Notably, the *region in text* baseline exhibits poor performance. Given the use of a powerful LLM in the pipeline, higher performance might be expected. Different variations were explored for this baseline, including not separating the coordinate digits or replacing coordinate digits with words, but performance did not improve. We hypothesize that the model fails to correctly map location information from the text to the image, which can be at least partly attributed to using a ViT to embed the image.

We provide qualitative example results in Fig. 4.4. The first column exemplifies cases where our method demonstrates robustness to subtle evidence (small biomarkers), correlations (surgical suture is usually close to the needle driver), and borderline cases (evidence close to the region). The second column highlights the weaknesses of *context only* when the context fails to provide enough evidence for the answer. Finally, the third column shows errors made by our model. Fig. 4.5 shows error maps by region location for the DME-VQA and INSEGCAT-VQA datasets and for the four strongest baselines. On the left side of the plot, the locations of actual positives and negatives are illustrated. For the INSEGCAT-VQA dataset, this visualization

Dataset	Method	Accuracy (%)	F1 score (%)
DME-VQA	No Mask	57.32	57.32
	Region in Text [129]	62.12	63.59
	Crop Region [149]	86.52	87.26
	Draw Region	86.86	86.85
	Context Only	88.07	88.45
	Ours	**90.30**	**90.22**
	LocAtt [159]*	84.2	85.79
RIS-VQA	No Mask	50.00	50.00
	Region in Text [129]	64.81	65.39
	Crop Region [149]	85.50	85.64
	Draw Region	91.30	91.43
	Context Only	91.77	91.81
	Ours	**92.60**	**92.54**
	LocAtt [159]*	82.73	86.15
INSEGCAT-VQA	No Mask	50.00	50.00
	Region in Text [129]	73.51	74.55
	Crop Region [149]	90.91	90.93
	Draw Region	95.44	95.43
	Context Only	95.19	95.17
	Ours	**95.51**	**95.47**
	LocAtt [159]*	88.13	90.14

TABLE 4.2: Accuracy and F1 score comparison to SOTA approaches on the DME-VQA, RIS-VQA and INSEGCAT-VQA datasets. For the DME-VQA dataset, only localized questions are considered (performance on other question types can be found in the supplementary materials). *This result corresponds to a different architecture, but we include it for completeness.

reveals a location bias that other baselines without access to the region or the context may be exploiting. Due to the nature of the images (cataract surgery) and questions, regions with positive answers tend to cluster in a specific area. This, coupled with the dissimilarity of objects mentioned in the questions, explains why a baseline like *crop region* achieves relatively high performance on this dataset compared to the other two datasets (see Table 4.2). Similarly, in the case of DME-VQA, it becomes evident that the lack of context in *crop region* results in lower sensitivity, highlighting the significance of context even when the isolated region should theoretically provide sufficient evidence. Fig. 4.5 also demonstrates that *draw region* and *context only* exhibit marked clusters of false positives and false negatives, potentially indicating the utilization of location biases. In contrast, our method produces a more evenly distributed location for both types of errors.

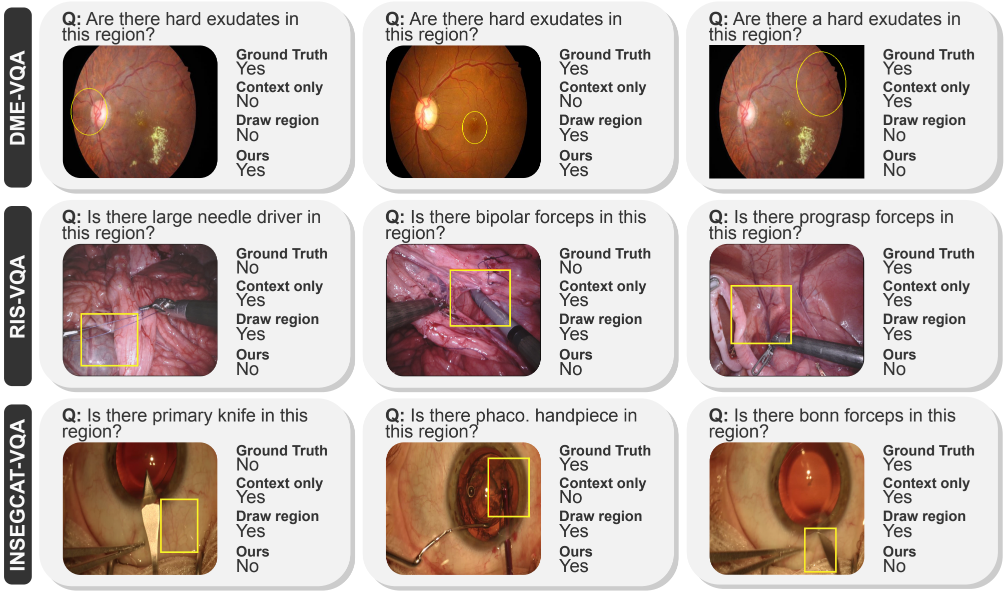


FIGURE 4.4: Qualitative examples on the DME-VQA (first row), RIS-VQA (second row), and INSEGCAT-VQA (third row) datasets. See Appendix B for additional examples.

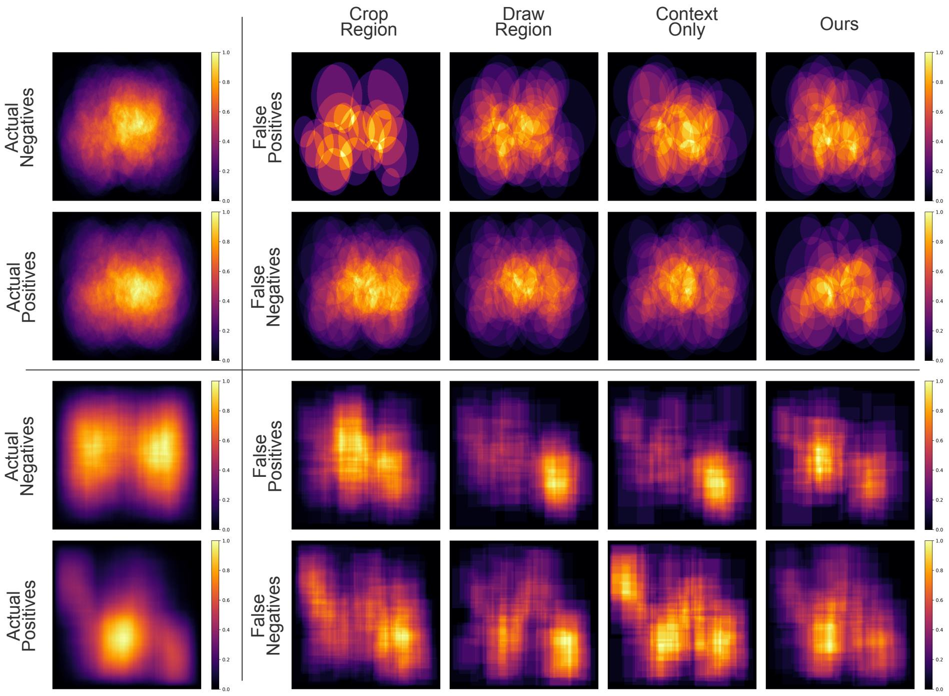


FIGURE 4.5: Error analysis by region location for the four strongest baselines. The maps are obtained by adding binary masks representing the regions for all QA pairs in each category and then normalizing. **Top:** DME-VQA dataset. **Bottom:** INSEGCAT-VQA dataset. The maps for RIS-VQA can be found in Appendix B.

4.4 Conclusion

In this work, we introduced a novel approach to enable localized questions in multimodal LLMs for the tasks of VQA. Our proposed approach involves the utilization of targeted visual prompting, granting the model access not only to the region and its context within the image but also to an isolated version of the region. By doing so, we allow two perspectives to be encoded in the prompt and allow more fine-grained information to be leveraged. Our approach demonstrates enhanced performance across all evaluated datasets compared to a variety of baselines. Analysis of the results highlights how biases in the datasets can be interpreted and qualitative examples shown depict failure modes of our method. Future works include extending the methodology to accommodate multiple images and enabling the use of comparison questions.

Enhancing Consistency in VQA Part II

5 Consistency-preserving Visual Question Answering in Medical Imaging

Since VQA models can be asked multiple questions about the same image, one important aspect of their behavior is what constraints there should be in the answers, given that the questions are related. This is, what level of agreement there should be in the answers so that these do not produce a contradiction. Most of the research in Med-VQA has been focused on improving architectures and working with limited data, while consistency has been overlooked.

In this work, we tackle the issue of inconsistency in Med-VQA by using a novel loss function term and corresponding training strategy that allows us to consider relations between question-answer pairs in the training process. Following previous approaches from natural images, we examine the case in which the relation between reasoning and perception questions is known. We evaluate our proposed approach on the task of DME staging from fundus images. Our experimental results show that our approach enhances not only the consistency of the model but also the overall performance.

Author Contribution This work was co-authored with Pablo Márquez-Neila and Raphael Sznitman. My contributions include the dataset creation, the formulation and implementation of the method, the experimental setup, result analysis and visualization, and the composition of the manuscript.

Publication This work is published in the Proceedings of the MICCAI 2022 conference [149].

License This work is published under the Springer License ↗. Reproduced with permission from Springer Nature.

5.1 Background and Previous Work

VQA models are neural networks that answer natural language questions about an image by interpreting the question and the image provided [29, 39, 115, 135]. Specifying questions using natural language gives VQA models great appeal, as the set of possible questions one can ask is enormous and does not need to be identical to the set of questions used to train the models. Due to these advantages, VQA models for medical applications have also been proposed [103, 104, 125, 127, 129, 130], whereby allowing clinicians to probe the model with subtle differentiating questions and contributing to build trust in predictions.

To date, much of the work in Med-VQA has focused on building more effective model architectures [104, 129, 130] or overcoming limitations in Med-VQA datasets [104, 127, 162, 173]. Yet a critical component of VQA is the notion of *consistency* in the answers produced by a model. Here, consistency refers to a model's capacity to produce answers that are not self-contradictory. For instance, the task of staging DME from color fundus photograph illustrated in Fig. 5.1 involves identifying *perception* elements in the image (*e.g.*, "are there hard exudates visible near the macula?") to infer a disease stage, which can be expressed as a *reasoning* question (*e.g.*, "what is the stage of disease?"). Ultimately, for any VQA model to be trustworthy, it should be able to answer these without contradicting itself (*i.e.*, answer that the image is healthy, but also identify hard exudates in the periphery of the eye).

Consistency in VQA has been been studied in the broader computer vision context [170, 175–178], where the relation between perception and reasoning questions is unconstrained. That is, the answers to perception questions do not necessarily imply any information with respect to the reasoning question and vice-versa. In these broad cases, some methods have modeled question implications [170, 177] or rephrased questions [178] by generating tailored question-answer pairs (*e.g.*, consistent data-augmentation). Alternatively, [176, 179, 180] used relations between questions to impose constraints in the VQA's embedding space. To avoid needing to know the relation between questions, [40] proposed to enforce consistency by making attention maps of reasoning and perception questions similar to one another. However, even though these approaches tackle unconstrained question relations, the ensuring of VQA models' consistency remains limited and often reduces the overall performance [40].

Instead, we propose a novel approach to enforce VQA consistency that is focused on cases where answers to the perception questions have explicit implications on reasoning question answers and vice-versa (*e.g.*, cancerous cells and severity of cancer found in H&E staining, or presence of hard exudates and DME staging). By focusing on this subset of question relations, our aim is to improve both the accuracy of our model and its consistency, without needing external data as in [162, 170, 175]. To do this, we allow questions to probe arbitrary image regions by masking irrelevant parts of the image and passing the masked image to the VQA model (see Fig. 5.1). To then enforce consistency, we propose a new loss function that penalizes incorrect perceptual predictions when reasoning ones are correct for a given image. To validate the impact of our approach, we test it in the context of DME staging and show

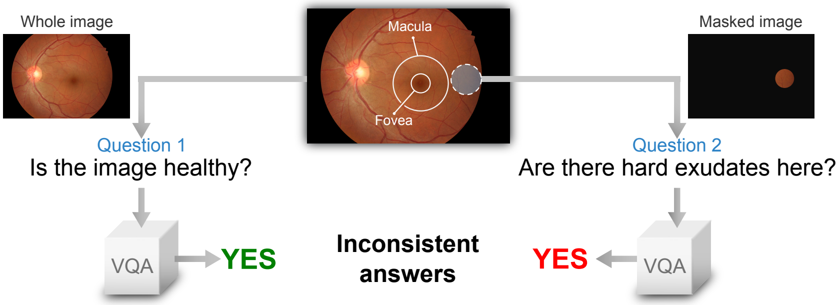


FIGURE 5.1: VQA inconsistency in Diabetic Macular Edema staging from fundus photograph. While the VQA model correctly answers “Is the image healthy?" (left), it incorrectly answers yes to “Are there hard exudates here?" for a specified retinal region (right).

that it outperforms state-of-the-art methods for consistency, without compromising overall performance accuracy.

5.2 Method

We present here our approach, which consists of using a simple VQA model with a training protocol that encourages consistency among pairs of perception and reasoning questions. Fig. 5.2 illustrates this VQA model and our training approach.

5.2.1 VQA Model

Following [181], our VQA model, $f : \mathcal{I} \times \mathcal{Q} \to \mathcal{P}(\mathcal{A})$, takes a tuple containing an image, $\mathbf{x}$, and a question, $\mathbf{q}$, to produce a distribution, $\mathbf{p} = f(\mathbf{x}, \mathbf{q})$, over a finite set of possible answers $\mathcal{A}$ (see Fig. 5.2(Top)). After encoding the inputs, the VQA model combines visual (v) and textual (q) features through an attention module (k) [182] that selects the visual features relevant to the question (v'). The final classifier receives a combination of the relevant features and the text features through a fusion module to predict the final distribution.

In some cases, questions may be asking about content related to specific regions of the image (*e.g.*, “are there hard exudates in this region?”). To process these cases, the input image is masked so that the visible area corresponds to the region mentioned in the question while the rest of the image is set to zero.

Training this model requires a dataset $\mathcal{T} = \{t^{(i)} = (\mathbf{x}^{(i)}, \mathbf{q}^{(i)}, a^{(i)})\}_{i=1}^{N} \subseteq \mathcal{I} \times \mathcal{Q} \times \mathcal{A}$ of images and questions annotated with their answers. The VQA loss is simply the cross-entropy between the predicted distribution and the real answer,

$$\ell_{\mathrm{VQA}}(\mathbf{p}, a) = H(\mathbf{p}, a) = -\log \mathbf{p}_a. \tag{5.1}$$

While this loss alone is sufficient to reach a reasonable performance, it ignores the potentially useful interactions that may exist among training questions.

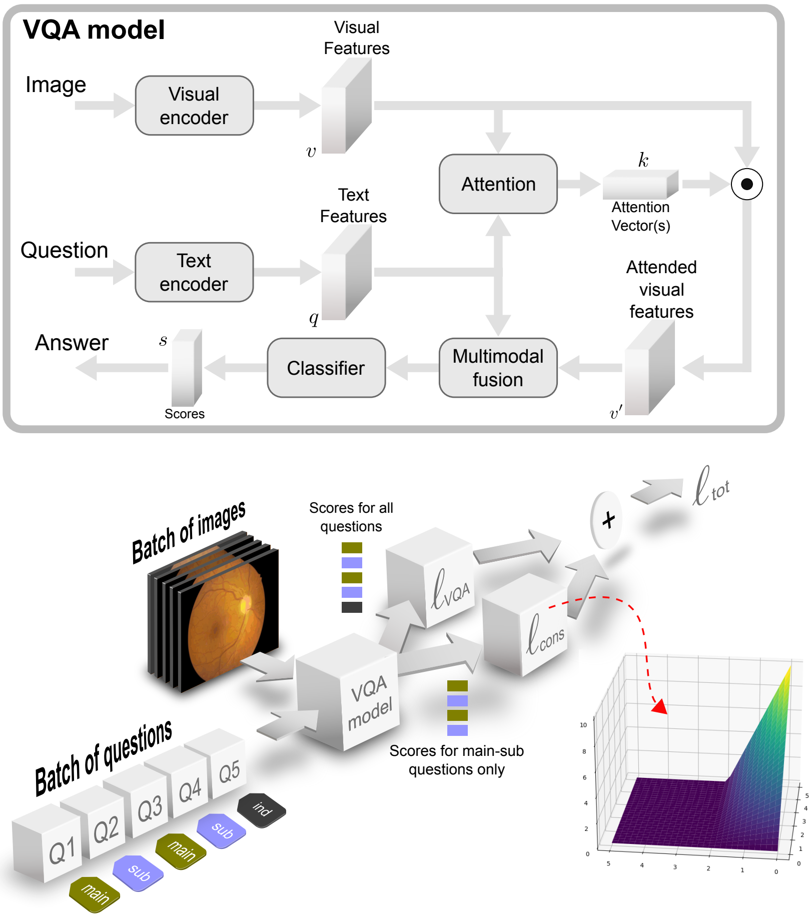


FIGURE 5.2: **Top:** VQA model architecture. **Bottom:** Visualization of the training process with the proposed loss. The total loss, ℓ_{tot}, is based on two terms: the conventional VQA loss, ℓ_{VQA} and our proposed consistency loss term, ℓ_{cons}. The latter is computed only for pairs of main (reasoning) and sub (perception) questions. Training mini-batches consist of main and sub-questions at the same time, whereby sub-questions can consider specific regions of the image. Unrelated questions (denoted with "ind") can also be included in training batches but do not contribute to ℓ_{cons}.

5.2.2 Consistency Loss

We aim to improve the quality of our VQA model by exploiting the relationships between reasoning and perception questions at training time. To this end, we augment the training dataset with an additional binary relation $\prec$ over the set of questions $\mathcal{Q}$. Two questions are related, $\mathbf{q}^{(i)} \prec \mathbf{q}^{(j)}$, if $\mathbf{q}^{(i)}$ is a perception question associated to the reasoning question $\mathbf{q}^{(j)}$. From hence on, we refer to perception questions as *sub-questions* and reasoning questions as *main questions.*

Following the terminology in [40], an inconsistency occurs when the VQA model infers the main question correctly but the sub-question incorrectly. Using the entropy as a measurement of incorrectness, we propose to impose the consistency at training time by penalizing the cases with high $H^{(i)} = H(\mathbf{p}^{(i)}, a^{(i)})$ and low $H^{(j)} = H(\mathbf{p}^{(j)}, a^{(j)})$ when $\mathbf{q}^{(i)} \prec \mathbf{q}^{(j)}$. To do this, we use an adapted hinge loss that disables the penalty when $H^{(j)}$ is larger than a threshold $\gamma > 0$, but otherwise penalizes large values of $H^{(i)}$,

$$\ell_{\text{cons}}(H^{(i)}, H^{(j)}) = H^{(i)} \max\{0, \gamma - H^{(j)}\}. \tag{5.2}$$

The final cost function then minimizes the expected value of the VQA loss (5.1) for the elements of the training dataset and the consistency loss (5.2) for the pairs of training samples with $\prec$-related questions,

$$\mathbb{E}_{t\sim\mathcal{T}}[\ell_{\text{VQA}}(\mathbf{p}, a)] + \lambda\mathbb{E}_{(t^{(i)},t^{(j)})\sim\mathcal{T}^2}[\ell_{\text{cons}}(H^{(i)}, H^{(j)}) \mid \mathbf{x}^{(i)} = \mathbf{x}^{(j)}, \mathbf{q}^{(i)} \prec \mathbf{q}^{(j)}], \tag{5.3}$$

where $\lambda > 0$ controls the relative strength of both losses and $\mathcal{T}^2$ is the Cartesian product of $\mathcal{T}$ with itself, that is, all pairs of training samples.

To train, this cost is iteratively minimized approximating the expectations with mini-batches. The two expectations of Eq. (5.3) suggest that two mini-batches are necessary at each iteration: one mini-batch sampled from $\mathcal{T}$ and a second mini-batch of $\prec$-related pairs sampled from $\mathcal{T}^2$. However, in practice a single mini-batch is sufficient as long as we ensure that it contains pairs of $\prec$-related questions. While this biased sampling could in turn bias the estimation of the first expectation, we did not observe a noticeable impact in our experiments. Fig. 5.2(Bottom) illustrates this training procedure.

5.3 Experiments and Results

5.3.1 DME Staging

DME staging from color fundus images involves grading images on a scale from 0 to 2, with 0 being healthy and 2 being severe (see Fig. 5.3). Differentiation between the grades relies on the presence of hard exudates present in different locations of the retina. Specifically, a grade of 0 implies that no hard exudates are present at all, a grade of 1 implies that hard exudates are located in the retina periphery (*i.e.*, outside a circular region centered at the fovea center with radius of one optic disc diameter), and a grade of 2 when hard exudates are in the macular region [158].

5.3.2 Dataset

To validate our method, we make use of two publicly available datasets: the Indian Diabetic Retinopathy Image Dataset (IDRiD) [156] and the e-Ophta dataset [183]. From the IDRiD dataset, we use images from the segmentation and grading tasks, which consist of 81 and 516 images, respectively. Images from the segmentation task include segmentation masks for hard exudates and images from the grading task only have the DME grade. On the other hand, the e-Ophta dataset comprises 47 images with segmentation of hard exudates and 35 images without lesions. Combining both datasets yields a dataset of 128 images with segmentation masks for hard exudates and 128 images without any lesions, plus 423 images for which only the DME risk grade is available.

In this context, we consider main questions to be those asking “What is the DME risk grade?" when considering the entire image. Sub-questions were then defined as questions asking about the presence of the hard exudates. For instance, as shown in Fig. 5.3(Right), “Are there hard exudates in this region?" where the region designated contains the macula. In practice, we set three types of sub-questions: “are there hard exudates in this image?", “are there hard exudates in the macula?" and “are there hard exudates in this region?". We refer to these three questions as *whole*, *macula* and *region* questions, respectively. For the region sub-questions, we consider circular regions that can be centered anywhere, or centered on the fovea, depending on availability of fovea center location annotations. As mentioned in Sec. 5.2, to answer questions about regions, images are masked so that only the region is visible.

The total number of question-answer pairs in our dataset consists of 9779 for training (4.4%

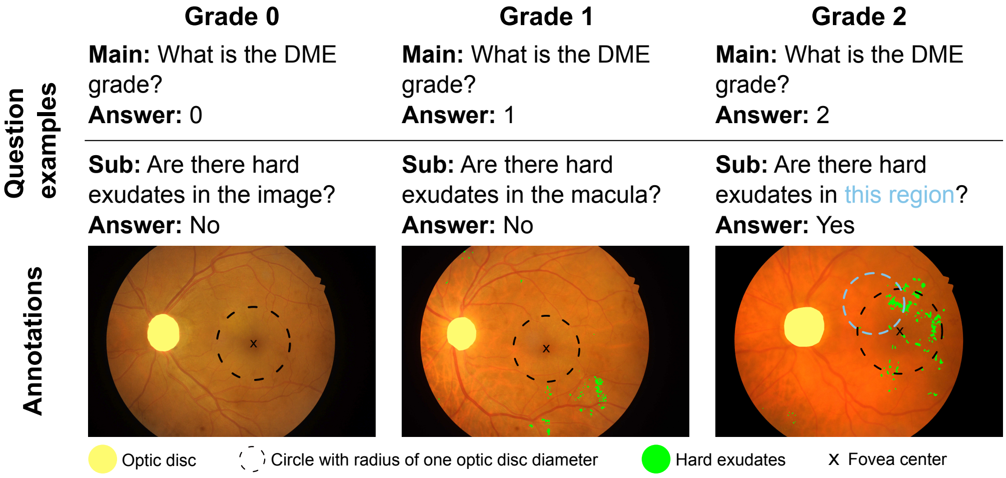

Grade 0
Grade 1
Grade 2
Question examples
Main: What is the DME grade?
Answer: 0
Main: What is the DME grade?
Answer: 1
Main: What is the DME grade?
Answer: 2
Sub: Are there hard exudates in the image?
Answer: No
Sub: Are there hard exudates in the macula?
Answer: No
Sub: Are there hard exudates in this region?
Answer: Yes
Annotations
Optic disc
Circle with radius of one optic disc diameter
Hard exudates
x Fovea center

FIGURE 5.3: DME risk grading. Grade 0 is assigned if there are no hard exudates present in the whole image. Grade 1 is assigned if there are hard exudates, but only located outside a circle centered at the fovea with radius of one optic disc diameter. Grade 2 is assigned if there are hard exudates located within the circle. Examples of main and sub-questions are provided for each grade.

main, 21.4% sub, 74.2% ind), 2380 for validation (4.5% main, 19.2% sub, 76.3% ind) and 1311 for testing (10% main, 46.1% sub, 43.9% ind), with images in the train, validation and test sets being mutually exclusive.

5.3.3 Experimental Setup

We compare our approach to a baseline model that does not use the proposed ℓ_{cons} loss, equivalent to setting $\lambda = 0$. In addition, we compare our method against the attention-matching method, SQuINT [40], as it is a state-of-the-art alternative to our approach that can be used with the same VQA model architecture.

Our VQA model uses an ImageNet-pretrained ResNet101 [54] with input image of 448 × 448 pixels and an embedding of 2048 dimensions for the image encoding. For text encoding, a single-layer LSTM [52] network processes the input question with word encoding of length 300 and produces a single question embedding of 1024 dimensions. The multi-glimpse attention mechanism [182] uses 2 glimpses and dropout rate 0.25, and the multimodal fusion stage uses standard concatenation. The final classifier is a multi-layer perceptron with hidden layer of 1024 dimensions and dropout rate of 0.25. Hyperparameters λ and γ were empirically adjusted to 0.5 and 1.0, respectively.

All experiments were implemented using PyTorch 1.10.1 and run on a Linux machine with an NVIDIA RTX 3090 graphic card using 16 GB of memory and 4 CPU cores. All methods use the weighted cross-entropy as the base VQA loss function. Batch size was set to 64, and we used Adam for optimization with a learning rate of 10^{-4}. Maximum epoch number was 100 and we use early stopping policy to prevent overfitting, with patience of 20 epochs [1].

We report accuracy and consistency [40] performances, using two different definitions of consistency for comparison. Consistency, C1, is the percentage of sub-questions that are answered correctly when the main question was answered correctly. Consistency, C2, is the percentage of main questions that are answered correctly when all corresponding sub-questions were answered correctly.

5.3.4 Results

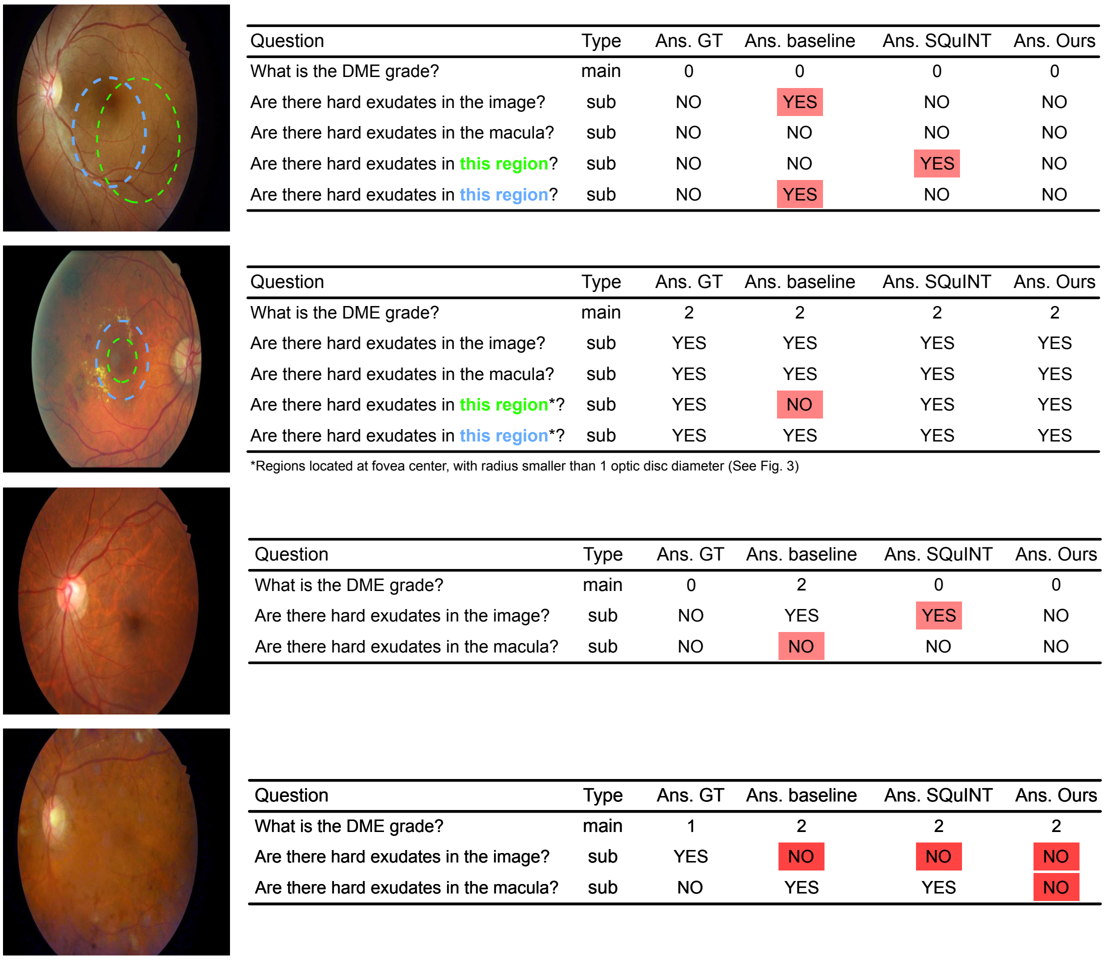
Table 5.1 shows the results. We compare these results to the case in which the value of λ is 0, which corresponds to the baseline in which no additional loss term is used. For each case, we present the overall accuracy and the accuracy for each type of question, as well as the consistency values. Fig. 5.4 illustrates the performance of each method with representative qualitative examples.

In general, we observe that our proposed approach yields increases in accuracy and consistency when compared to both the baseline and SQuINT. Importantly, this increase in

[1] Our code and data are available at https://github.com/sergiotasconmorales/consistency_vqa

Case	Accuracy					Consistency	
	overall	grade	whole	macula	region	C1	C2
Baseline (no att.)	77.54	73.59	81.37	83.37	76.66	81.70	91.86
Baseline (att.)	81.46	80.23	83.13	**87.18**	80.58	89.21	96.92
Baseline (att.) + SQuINT [40]	80.58	77.48	82.82	85.34	80.02	88.17	94.62
Baseline (att.) + Ours ($\lambda = 0.5, \gamma = 1$)	**83.49**	**80.69**	**84.96**	**87.18**	**83.16**	**94.20**	**98.12**

TABLE 5.1: Average test accuracy and consistency values for the different models. Results shown are averaged over 10 models trained with different seeds. Accuracy values are presented for all questions (overall), for main questions (grade) and for sub-questions (whole, macula and region). Both measures of consistency are shown as well.



Question	Type	Ans. GT	Ans. baseline	Ans. SQuINT	Ans. Ours
What is the DME grade?	main	0	0	0	0
Are there hard exudates in the image?	sub	NO	YES	NO	NO
Are there hard exudates in the macula?	sub	NO	NO	NO	NO
Are there hard exudates in this region?	sub	NO	NO	YES	NO
Are there hard exudates in this region?	sub	NO	YES	NO	NO

Question	Type	Ans. GT	Ans. baseline	Ans. SQuINT	Ans. Ours
What is the DME grade?	main	2	2	2	2
Are there hard exudates in the image?	sub	YES	YES	YES	YES
Are there hard exudates in the macula?	sub	YES	YES	YES	YES
Are there hard exudates in this region*?	sub	YES	NO	YES	YES
Are there hard exudates in this region*?	sub	YES	YES	YES	YES

*Regions located at fovea center, with radius smaller than 1 optic disc diameter (See Fig. 3)

Question	Type	Ans. GT	Ans. baseline	Ans. SQuINT	Ans. Ours
What is the DME grade?	main	0	2	0	0
Are there hard exudates in the image?	sub	NO	YES	YES	NO
Are there hard exudates in the macula?	sub	NO	NO	NO	NO

Question	Type	Ans. GT	Ans. baseline	Ans. SQuINT	Ans. Ours
What is the DME grade?	main	1	2	2	2
Are there hard exudates in the image?	sub	YES	NO	NO	NO
Are there hard exudates in the macula?	sub	NO	YES	YES	NO

FIGURE 5.4: Qualitative examples from the test set. Inconsistent sub-answers are highlighted in red. Additional examples are shown in Appendix C.

consistency is not at the expense of overall accuracy. Specifically, this indicates that our loss term causes the model to be correct about sub-questions when it is correct about main ques-

λ	γ	Accuracy					Consistency	
		overall	grade	whole	macula	region	C1	C2
0	-	81.46	80.23	83.13	87.18	80.58	89.21	96.92
0.2	0.5	82.01	80.38	83.59	86.56	81.36	90.93	97.38
0.2	1	82.65	79.77	83.97	86.64	82.30	93.22	97.51
0.2	1.5	83.05	81.22	84.27	87.33	82.53	93.23	97.56
0.3	0.5	82.34	79.92	83.59	87.71	81.74	92.32	97.31
0.3	1	83.27	80.53	84.58	87.25	82.91	94.01	98.10
0.3	1.5	83.28	80.84	84.43	87.48	82.86	93.28	98.29
0.4	0.5	82.87	80.69	84.89	87.02	82.30	92.66	96.66
0.4	1	82.97	80.15	83.97	86.72	82.69	93.91	98.23
0.4	1.5	83.33	80.08	84.20	86.87	83.17	93.96	97.77
0.5	0.5	82.54	81.07	83.66	88.02	81.81	91.87	97.73
0.5	1	83.49	80.69	84.96	87.18	83.16	94.20	98.12
0.5	1.5	83.25	79.92	84.58	86.95	83.01	94.20	98.12

TABLE 5.2: Average test accuracy and consistency values for different values of the parameters λ and γ. The first row ($\lambda = 0$) corresponds to no consistency enhancement method.

tions. The observed increase in accuracy also indicates that our approach is not synthetically increasing consistency by reducing the number of correct answers on main questions [40]. We note that SQuINT results in a reduction in accuracy and consistency, which can be partially explained by the presence of region questions that are not associated to any main question. These questions, which exceed the number of main questions, may affect the constraint in the learned attention maps.

Table 5.2 shows the effect of λ and γ on the performance metrics. As expected, we notice that when λ increases, the consistency of our approach increases as well and will occasionally deteriorate overall accuracy. The impact of γ however is less evident, as no clear trend is visible. This would imply that the exact parameter value used is moderately critical to performances.

5.4 Conclusion

In this work, we presented a novel method for improving consistency in VQA models in cases where answers to sub-questions imply those of main questions and vice-versa. By using a tailored training procedure and loss function that measures the level of inconsistency, we show on the application of DME staging, that our approach provides important improvements in both VQA accuracy and consistency. In addition, we show that our method's hyperparameters are relatively insensitive to model performance. In the future, we plan to investigate how this approach can be extended to the broader case of unconstrained question relations.

6 Logical Implications for Visual Question Answering Consistency

The previous chapter presented a method to encourage a VQA method to be more consistent by considering pairs of reasoning and perception questions and their relationship. While this consideration is useful for imposing a more consistent behavior, the relation of main and sub (or reasoning and perception) between QA pairs requires assumptions about the nature of the questions. We observe that a more general definition of consistency is required, and explore, from the perspective of logic, the possible relations that can exist between the propositions that the QA pairs represent.

The present work presents a more general framework for consistency enforcement and assessment. We make use of concepts from logic in order to establish a more robust definition of inconsistency. Then, we encourage the model to provide more consistent answers by integrating annotations about logical relations between pairs of propositions into the training process. Since these annotations are not commonly included in VQA datasets, we propose an auxiliary method to predict them. We evaluate our method on multiple architectures and across different datasets with both natural and medical images, showing that our consistency framework improves the overall performance of the model while reducing inconsistencies.

Author Contribution The co-authors of this work are Pablo Márquez-Neila and Raphael Sznitman. My contributions are as follows: Data annotation, dataset creation, methodology and implementation thereof, experimental design, result analysis, and visualization, as well as the writing of the manuscript.

Publication This work is published in the Proceedings of the CVPR 2023 conference [184].

6.1 Background and previous work

6.1.1 Background

VQA models have drawn recent interest in the computer vision community as they allow text queries to question image content. This has given way to a number of novel applications in the space of model reasoning [185–188], medical diagnosis [105, 126, 127, 129] and counterfactual learning [189–191]. With the ability to combine language and image information in a common model, it is unsurprising to see a growing use of VQA methods.

Despite this recent progress, however, a number of important challenges remain when making VQAs more proficient. For one, it remains extremely challenging to build VQA datasets that are void of bias. Yet this is critical to ensure subsequent models are not learning spurious correlations or shortcuts [118]. This is particularly daunting in applications where domain knowledge plays an important role (*e.g.*, medicine [147, 153, 164]). Alternatively, ensuring that responses of a VQA are coherent, or *consistent*, is paramount as well. That is, VQA models that answer differently about similar content in a given image imply inconsistencies in how the model interprets the inputs. A number of recent methods have attempted to address this using logic-based approaches [176], rephrasing [178], question generation [170, 175, 177] and regularizing using consistency constraints [149]. In this work, we follow this line of research and look to yield more reliable VQA models.

We wish to ensure that VQA models are consistent in answering questions about images. This implies that if multiple questions are asked about the same image, the model's answers

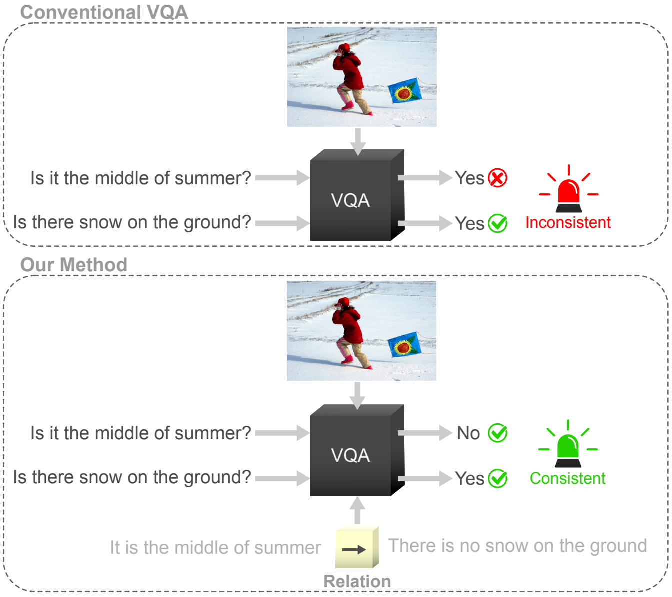

Conventional VQA
Is it the middle of summer?
Is there snow on the ground?
VQA
Yes
Yes
Inconsistent
Our Method
Is it the middle of summer?
Is there snow on the ground?
VQA
No
Yes
Consistent
It is the middle of summer
→
There is no snow on the ground
Relation

FIGURE 6.1: **Top:** Conventional VQA models tend to produce inconsistent answers as a consequence of not considering the relations between question and answer pairs. **Bottom:** Our method learns the logical relation between question and answer pairs to improve consistency.

should not contradict themselves. For instance, if one question about the image in Fig. 6.1 asks “Is there snow on the ground?", then the answer inferred should be consistent with that of the question “Is it the middle of summer?" As noted in [40], such question pairs involve reasoning and perception, and consequentially lead the authors to define inconsistency when the reasoning and perception questions are answered correctly and incorrectly, respectively. Along this line, [149] uses a similar definition of inconsistency to regularize a VQA model meant to answer medical diagnosis questions that are hierarchical in nature. What is critical in both cases, however, is that the consistency of the VQA model depends explicitly on its answers, as well as the question and true answer. This hinges on the assumption that perception questions are sufficient to answer reasoning questions. Yet, for any question pair, this may not be the case. As such, the current definition of consistency (or inconsistency) has been highly limited and does not truly reflect how VQAs should behave.

To address the need to have self-consistent VQA models, we propose a novel training strategy that relies on logical relations. To do so, we re-frame question-answer (QA) pairs as propositions and consider the relational construct between pairs of propositions. This construct allows us to properly categorize pairs of propositions in terms of their logical relations. From this, we introduce a novel loss function that explicitly leverages the logical relations between pairs of questions and answers in order to enforce that VQA models be self-consistent. However, datasets typically do not contain relational information about QA pairs, and collecting this would be extremely laborious and difficult. To overcome this, we propose to train a dedicated language model that infers logical relations between propositions. Our experiments show that we can effectively infer logical relations from propositions and use them in our loss function to train VQA models that improve state-of-the-art methods via consistency. We demonstrate this over two different VQA datasets, against different consistency methods, and with different VQA model architectures.

6.1.2 Previous Work

Since its initial presentation in Antol et al. [39], VQA has thoroughly advanced. Initial developments focused on multimodal fusion modules, which combine visual and text embeddings [186, 192]. From basic concatenation and summation [39] to more complex fusion mechanisms that benefit from projecting the embeddings to different spaces, numerous approaches have been proposed [106, 110, 166]. The addition of attention mechanisms [107, 186, 192] and subsequently transformer architectures [27] has also contributed to the creation of transformer-based VLM, such as LXMERT, which have shown state-of-the-art performances [115].

More recently, methods have proposed to improve other aspects of VQA, including avoiding shortcut learning and biases [193, 194], improving 3D spatial reasoning [195], Out-Of-Distribution (OOD) generalization [118, 196], improving transformer-based vision-language models [197, 198], external knowledge integration [199, 200] and model evaluation with visual and/or textual perturbations [201, 202]. With the awareness of bias in VQA training data, some works have also addressed building better datasets (*e.g.*, VQAv2.0 [135], VQA-CP [136], CLEVR [145] and GCP [29]).

Furthermore, these developments have now given rise to VQA methods in specific domains. For instance, the VizWiz challenge [34, 35, 203] aims at creating VQA models that can help visually impaired persons with routine daily tasks, while there is a growing number of Med-VQA works with direct medicine applications [105, 126, 127, 129].

Consistency in VQA

Consistency in VQA can be defined as the ability of a model to produce answers that are not contradictory. This is, given a pair of questions about an image, the answers predicted by a VQA model should not be contrary (*e.g.*answering “Yes" to “Is it the middle of summer?" and “Winter" to “What season is it?"). Due to its significance in reasoning, consistency in VQA has become a focus of study in recent years [40, 170, 176, 178, 188]. Some of the first approaches for consistency enhancement focused on creating re-phrasings of questions, either by dataset design or at training time [178]. Along this line, entailed questions were proposed [170, 176], such that a question generation module was integrated into a VQA model [175, 177], used as a benchmarking method to evaluate consistency [180] or as a rule-based data-augmentation technique [170]. Other approaches tried to shape the embedding space by imposing constraints in the learned representations [179] and by imposing similarities between the attention maps of pairs of questions [40]. Another work [149] assumed entailment relations between pairs of questions to regularize training. A more recent approach attempts to improve consistency by using graph neural networks to simulate a dialog in the learning process [188].

While these approaches show benefits in some cases, they typically only consider that a subset of logical relationships exists between pairs of question-answers or assume that a single relation holds for all QA pairs. Though true in the case of re-phrasings, other question generation approaches cannot guarantee that the produced questions preserve unique relations or that grammatical structure remains valid. Consequently, these methods often rely on metrics that either over or under-estimate consistency by relying on these assumptions. In the present work, we propose a strategy to alleviate these limitations by considering all logical relations between pairs of questions and answers.

Entailment Prediction

Natural Language Inference (NLI), or Recognizing Textual Entailment (RTE), is the task of predicting how two input sentences (namely *premise* and *hypothesis*) are related, according to three pre-established categories: entailment, contradiction and neutrality [204]. For example, if the premise is “A soccer game with multiple males playing" and the hypothesis is “Some men are playing a sport," then the predicted relation should be an entailment, because the hypothesis logically follows from the premise. Several benchmarking datasets (*e.g.*, SNLI [205], MultiNLI [206], SuperGLUE [207], WIKI-FACTCHECK [208] and ANLI [209]) have contributed to the adaption of general-purpose transformer-based models like Bidirectional Encoder Representations from Transformers (BERT) [48], RoBERTa [210] and DeBERTa [211] for this task. In this work, we will leverage these recent developments to build a model capable of inferring relations between propositions.

6.2 Method

Given an image $\mathbf{x} \in \mathcal{I}$, a question $\mathbf{q} \in \mathcal{Q}$ about the image and a set $\mathcal{A} = \{a_1, \ldots, a_K\}$ of possible answers to choose from, a VQA model is expected to infer the answer $\hat{a} \in \mathcal{A}$ that matches the true answer a^*. This can be formulated as,

$$\hat{a} = \underset{a \in \mathcal{A}}{\arg\max}\, p(a|\mathbf{x}, \mathbf{q}; \boldsymbol{\theta}), \tag{6.1}$$

where $\boldsymbol{\theta}$ represents the parameters of the VQA model.

In this context, we observe that two QA pairs $(\mathbf{q}_i, a_i)$ and $(\mathbf{q}_j, a_j)$ for the same image $\mathbf{x}$ can have different kinds of logical relations. In the simplest case, the two pairs may be unrelated, as with the pairs (“Is it nighttime?”, “Yes”) and (“Is there a bench in the image?”, “No”). Knowing that one of the pairs is true gives no information about the truth value of the other.

On the other hand, two pairs may be related by a logical implication, as in the pairs (“Is the horse brown?", “No") and (“What is the color of the horse?", “White"). Knowing that the second pair is true implies that the first pair must be true as well. Conversely, if the first pair is false (*the horse is brown*), it implies that the second pair must also be false. In this case, the first pair is a necessary condition for the second one or, equivalently, the second pair is a sufficient condition for the first one.

Finally, it can be that two QA pairs are related by a double logical implication, as with the pairs (“Is this a vegetarian pizza?”, “Yes”) and (“Does the pizza have meat on it?”, “No”). The veracity of the former implies the veracity of the latter, but the veracity of the latter also implies the veracity of the former. In this case, each pair is simultaneously a necessary and sufficient condition for the other pair, and both pairs are then equivalent.

Note that the logical implication existing between two QA pairs is an intrinsic property of the QA pairs, and does not depend on the correctness of the predictions coming from a VQA model. If a VQA model considers a sufficient condition true and a necessary condition false, it is incurring an *inconsistency* regardless of the correctness of its predictions.

Since logical implications are the basis of reasoning, we propose to explicitly use them when training a VQA model to reduce its inconsistent predictions. Unfortunately, doing so requires overcoming two important challenges: (1) a strategy is needed to train VQA models with logical relations that leverage consistency in a purposeful manner. Until now, no such approach has been proposed; (2) VQA datasets do not typically contain logical relations between pairs of QA. Acquiring these manually would, however, be both time-consuming and difficult.

We address these challenges in this work by formalizing the idea of consistency and treating QA pairs as logical propositions from which relations can comprehensively be defined. Using this formalism, we first propose a strategy to solve (1) and train a VQA model more effectively using logical relations and the consistency they provide (Sec. 6.2.2). We then show in Sec. 6.2.3 how we infer relations between pairs of propositions, thereby allowing standard VQA datasets to be augmented with logical relations.

6.2.1 Consistency Formulation

We begin by observing that QA pairs $(\mathbf{q}, a)$ can be considered and treated as logical propositions. For instance, the QA (“Is it winter?", “Yes") can be converted to “It is winter," which is a logical proposition that can be evaluated as *true* or *false* (*i.e.*, its *truth value*). Doing so allows us to use a broad definition of consistency, namely one that establishes that two propositions are inconsistent if both cannot be true at the same time [212]. In the context of this work, we assume the truth value of a proposition $(\mathbf{q}, a)$ is determined by an agent (either a human annotator or the VQA model) after observing the information contained in an image $\mathbf{x}$.

Let $\mathcal{D} = \mathcal{I} \times \mathcal{Q} \times \mathcal{A}$ be a VQA dataset that contains triplets $(\mathbf{x}^{(n)}, \mathbf{q}_i^{(n)}, a_i^{(n)})$, where $\mathbf{x}^{(n)}$ is the n-th image and $(\mathbf{q}_i^{(n)}, a_i^{(n)})$ is the i-th question-answer pair about $\mathbf{x}^{(n)}$. In the following, we omit the index n for succinctness. For a given image $\mathbf{x}$, we can consider a pair of related question-answers as $(\mathbf{q}_i, a_i)$ and $(\mathbf{q}_j, a_j)$ as a pair of propositions. Following propositional logic notation, if both propositions are related in such a way that $(\mathbf{q}_i, a_i)$ is a sufficient condition for the necessary condition $(\mathbf{q}_j, a_j)$, we write that $(\mathbf{q}_i, a_i) \rightarrow (\mathbf{q}_j, a_j)$. For convenience, this arrow notation can be adapted to indicate different orderings between the necessary and sufficient conditions:

- $(\mathbf{q}_i, a_i) \leftarrow (\mathbf{q}_j, a_j)$ if the proposition $(\mathbf{q}_i, a_i)$ is a necessary condition for $(\mathbf{q}_j, a_j)$.
- $(\mathbf{q}_i, a_i) \leftrightarrow (\mathbf{q}_j, a_j)$ if the propositions $(\mathbf{q}_i, a_i)$ and $(\mathbf{q}_j, a_j)$ are equivalent, *i.e.*, both are simultaneously necessary and sufficient. Note that this is just notational convenience for the double implication $(\mathbf{q}_i, a_i) \rightarrow (\mathbf{q}_j, a_j) \wedge (\mathbf{q}_j, a_j) \rightarrow (\mathbf{q}_i, a_i)$, and in the following derivations the double arrow will be always considered as two independent arrows.
- Finally, we will write $(\mathbf{q}_i, a_i) - (\mathbf{q}_j, a_j)$ if the propositions $(\mathbf{q}_i, a_i)$ and $(\mathbf{q}_j, a_j)$ are not related.

If a VQA model is asked questions $\mathbf{q}_i$ and $\mathbf{q}_j$ about an image $\mathbf{x}$ and there exists a relation $(\mathbf{q}_i, a_i) \rightarrow (\mathbf{q}_j, a_j)$, the answers of the model will be inconsistent whenever it provides answers $\hat{a}_i = a_i$ and $\hat{a}_j \neq a_j$ (*i.e.*, the model evaluates the first proposition as true and the second proposition as false). More generally, for a pair of necessary and sufficient conditions, the agent will be inconsistent if it evaluates the necessary condition as false and the sufficient condition as true [212]. In what follows, we exploit these ideas to quantify model inconsistencies in our experiments and to develop a new loss function that encourages logically consistent VQA models.

6.2.2 Logical Implication Consistency Loss

The core aim of our method is to encourage the VQA model to avoid inconsistent answers. When training, assume that the model receives an image $\mathbf{x}$ from $\mathcal{D}$ and two associated propositions $(\mathbf{q}_1, a_1)$ and $(\mathbf{q}_2, a_2)$ that are related by a logical implication $(\mathbf{q}_1, a_1) \rightarrow (\mathbf{q}_2, a_2)$. We define,

$$\pi_i = \pi\big((\mathbf{q}_i, a_i), \mathbf{x}\big) = p(a_i \mid \mathbf{x}, \mathbf{q}_i; \boldsymbol{\theta}), \tag{6.2}$$

as the probability assigned by the VQA model that the proposition $(\mathbf{q}, a)$ is true for the image $\mathbf{x}$. The model has a high probability of incurring an inconsistency if it simultaneously gives a high probability π_1 to the sufficient condition and a low probability π_2 to the necessary condition.

We thus define our consistency loss as a function,

$$\ell_{\text{cons}}(\mathbf{x}, (\mathbf{q}_1, a_1), (\mathbf{q}_2, a_2)) = -(1 - \pi_2)\log(1 - \pi_1) - \pi_1 \log(\pi_2), \tag{6.3}$$

that takes an image and a pair of sufficient and necessary propositions, and penalizes predictions with a high probability of inconsistency. As illustrated in Fig. 6.2, ℓ_{cons} is designed to produce maximum penalties when $\pi_1 = 1$ and $\pi_2 < 1$ (*i.e.*, when the sufficient condition is absolutely certain but the necessary condition is not), and when $\pi_2 = 0$ and $\pi_1 > 0$ (*i.e.*, when the necessary condition can never be true but the sufficient condition can be true). At the same time, ℓ_{cons} produces minimum penalties when either $\pi_1 = 0$ or $\pi_2 = 1$, as no inconsistency is possible when the sufficient condition is false or when the necessary condition is true. Interestingly, despite its resemblance, ℓ_{cons} is not a cross-entropy, as it is not an expectation over a probability distribution.

Our final loss is then a linear combination of the consistency loss and the cross-entropy loss ℓ_{VQA} typically used to train VQA models. Training with this loss then optimizes,

$$\min_{\theta} \mathbb{E}_{\mathcal{D}}[\ell_{\text{VQA}}] + \lambda \mathbb{E}_{\substack{((\mathbf{x}_i, \mathbf{q}_i, a_i), (\mathbf{x}_j, \mathbf{q}_j, a_j)) \sim \mathcal{D}^2 \\ \mathbf{x}_i = \mathbf{x}_j, (\mathbf{q}_i, a_i) \rightarrow (\mathbf{q}_j, a_j)}}[\ell_{\text{cons}}], \tag{6.4}$$

where the first expectation is taken over the elements of the training set $\mathcal{D}$ and the second expectation is taken over all pairs of necessary and sufficient propositions from $\mathcal{D}$ defined for the same image. In practice, we follow the sampling procedure described in [40, 149], where mini-batches contain pairs of related questions. The hyperparameter λ controls the relative strength between the VQA loss and the consistency term.

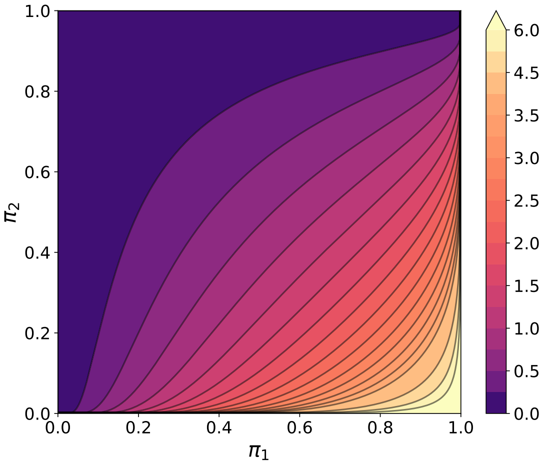


FIGURE 6.2: Consistency loss ℓ_{cons} as a function of the estimated probabilities for the sufficient, π_1, and necessary, π_2, conditions. Note that the loss diverges to ∞ when $\pi_1 = 1, \pi_2 < 1$ and when $\pi_1 > 0, \pi_2 = 0$.

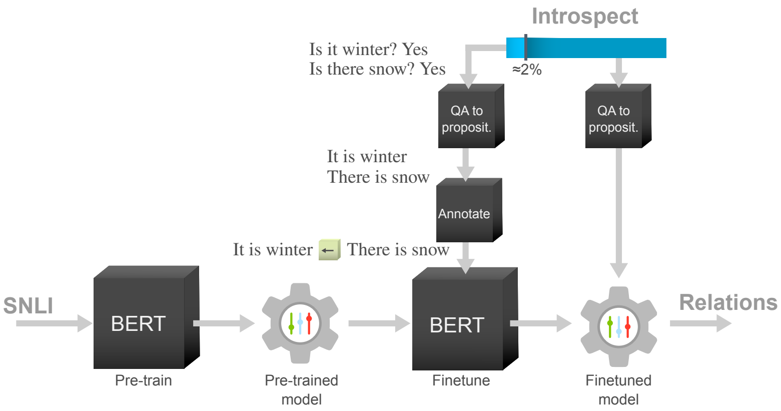


FIGURE 6.3: LI-MOD: Approach to predict logical relations between pairs of propositions. A BERT-based NLP model is first pre-trained on the SNLI dataset [205] to solve a Natural Language Inference task and subsequently fine-tuned with annotated pairs from a subset of Introspect dataset [40]. The resulting model is used to predict the relations of the remaining part of the dataset.

6.2.3 Inferring Implications

By and large, VQA datasets do not include annotations with logical relations between question-answers pairs, which makes training a VQA with ℓ_{cons} infeasible. To overcome this, we propose to train a language model to predict logical implications directly and use these predictions instead. We achieve this in two phases illustrated in Fig. 6.3 and refer to our approach as the Logical-Implication model (LI-MOD).

First, we pre-train BERT [48] on the task of Natural Language Inference using the SNLI dataset [205], which consists of pairs of sentences with annotations of entailment, contradiction or neutrality. In this task, given two sentences, a language model must predict one of the mentioned categories. While these categories do not exactly match the logical implication relevant to our objective, they can be derived from the entailment category. To this end, given two propositions $(\mathbf{q}_i, a_i)$ and $(\mathbf{q}_j, a_j)$, we evaluate them using the finetuned NLI model in the order $(\mathbf{q}_i, a_i), (\mathbf{q}_j, a_j)$, and then repeat the evaluation by inverting the order, to evaluate possible equivalences or inverted relations. If the relation is predicted as neutral in both passes, the pair is considered to be unrelated.

Then, we finetune the NLI model on a sub-set of annotated pairs from the VQA dataset Introspect [40]. In practice, we use a subset of binary QA pairs that were manually annotated with logical implications. Even though the relation need not be limited to binary questions (*i.e.*, yes/no questions), we chose to do so because the relation annotation is simpler than for open-ended questions. Since BERT expects sentences and not QA pairs, these were first converted into propositions using Parts Of Speech (POS) tagging [213] and simple rules that apply to binary questions (*e.g.*, to convert “Is it winter?," “Yes" we invert the first two words of the question and remove the question mark). After finetuning the model, the relations were predicted for the remaining part of the dataset. Further implementation details on this are given in 6.3.3.

6.3 Experiments and Results

We evaluate our proposed consistency loss function on different datasets and using a variety of VQA models.

6.3.1 Datasets

Introspect [40]

Contains perception questions (or sub-questions) created by annotators for a subset of reasoning questions (or main questions) of the VQA v1.0 and v2.0 datasets [39, 135]. It contains 27,441 reasoning questions with 79,905 sub-questions in its training set and 15,448 reasoning questions with 52,573 sub-questions for validation. For images that have the same sub-question repeated multiple times, we remove duplicates in the sub-questions for every image in both the train and validation sets.

DME Dataset [149]

Consists of retinal fundus images for the task of DME staging. It contains 9,779 QA pairs for training, 2,380 QA pairs for validation and 1,311 QA pairs for testing. There are three types of questions in the dataset: main, sub, and independent questions. Main questions ask about diagnosis information (*i.e.*the stage of the disease) and sub-questions ask about the presence and location of biomarkers. Sub-questions are further subdivided into grade questions, questions about the whole image, questions about a region of the eye called macula, and questions about random regions in the image. To enable questions about image regions, we follow the procedure described in [149], whereby only the relevant region is shown to the model.

6.3.2 Baseline Methods and Base Models

We consider 3 different consistency enhancement baseline methods. To ensure fair comparisons, all methods use the same VQA base models and only differ in the consistency method used. These consist in:

- *None:* Indicating that no consistency preserving method is used with the VQA model. This corresponds to the case where $\lambda = 0$.

- *SQuINT [40]:* Optimizes consistency by maximizing the similarity between the attention maps of pairs of questions. As such, it requires a VQA model that uses guided attention.

- *CP-VQA [149]:* Assumes entailment relations and uses a regularizer to improve consistency.

VQA Architectures

We show experiments using three VQA models depending on the dataset used. For experiments on Introspect, we make use of the BAN model [107], as its structure with guided attention allows the use of SQuINT. In addition, we evaluate the vision-language architecture LXMERT [115] on this dataset to evaluate improvement in state-of-the-art, transformer-based VQA models. For experiments on the DME dataset, we use the base model described in [149], which we denote by MVQA.

6.3.3 Implementation Details

LI-Model

We first pre-train BERT on SNLI for 5 epochs until it reaches a maximum accuracy of 84.32% on that dataset. For this pre-training stage, we initialize BERT with the *bert-base-uncased* weights and use a batch size of 16. We use a weight decay rate of 0.01 and the AdamW optimizer with a learning rate of $2 \cdot 10^{-5}$. The same setup was kept to finetune the model on a subset of 2'000 pairs of propositions from Introspect which were manually annotated (distribution of labels being: $\leftarrow 60\%, \leftrightarrow 17\%, -12\%, \rightarrow 11\%$), and an additional 500 pairs were annotated for validation. Notice that LI-MOD is only necessary for the Introspect dataset since, for the DME dataset, the implications annotations are available.

VQA Models

For our base models, we use the official and publicly available implementations (BAN [214], LXMERT [115] and MVQA [149]) with default configurations. We re-implemented SQuINT [40] and used the provided implementation of CP-VQA [149], reporting the best results, which were obtained with $\lambda = 0.1, \gamma = 0.5$ for BAN and $\lambda = 0.5, \gamma = 1$ for MVQA (parameters refer to original implementations). For SQuINT, we set the gain of the attention map similarity term to 0.5 for BAN and 1.0 for MVQA. For Introspect, we train 5 models with different seeds for each parameter set and for DME, we train 10 models with different seeds. To train LXMERT, BAN and MVQA, we use batch sizes of 32, 64 and 128, respectively. Regarding the VQA cross-entropy loss, we follow the original implementations and use soft scores for the answers in LXMERT and categorical answers for BAN and MVQA.

6.3.4 Quantifying Consistency

Given a test set $\mathcal{T} = \{t_n\}_{n=1}^{|\mathcal{T}|}$, where $t_n = (\mathbf{x}, \mathbf{q}, a)$ is a test sample triplet, we wish to measure the level of consistency of a VQA model p. To this end, we define the set of implications $G(\mathcal{T}) \subset \mathcal{T}^2$ as the collection of all pairs of test samples $((\mathbf{x}_i, \mathbf{q}_i, a_i), (\mathbf{x}_j, \mathbf{q}_j, a_j))$ for which $(\mathbf{q}_i, a_i) \rightarrow (\mathbf{q}_j, a_j)$ and $\mathbf{x}_i = \mathbf{x}_j$, and the set of inconsistencies $I_p(\mathcal{T})$ produced by the VQA model as the subset of $G(\mathcal{T})$ that contains the pairs for which the model evaluated the sufficient condition as true and the necessary condition as false,

$$I_p(\mathcal{T}) = \{(t_i, t_j) \in G(\mathcal{T}) \mid e_p((\mathbf{q}_i, a_i), \mathbf{x}) \wedge \neg e_p((\mathbf{q}_j, a_j), \mathbf{x})\}.$$

The function e_p returns the truth value of the proposition $(\mathbf{q}, a)$ for image $\mathbf{x}$ evaluated by the VQA model p,

$$e_p((\mathbf{q}, a), \mathbf{x}) = [\hat{a} = a], \tag{6.5}$$

where $\hat{a}$ is the answer of maximum probability following Eq. (6.1). In other words, e_p returns whether the estimated answer for question $\mathbf{q}$ matches the answer of the proposition a. Finally, the consistency ratio c for model p on the test set $\mathcal{T}$ is the proportion of implications in $G(\mathcal{T})$ that did not lead to an inconsistency,

$$c_p(\mathcal{T}) = 1 - \frac{|I_p(\mathcal{T})|}{|G(\mathcal{T})|}. \tag{6.6}$$

6.3.5 Results

Performance Comparison

For both datasets, we first compare the performance of our method against the baseline consistency methods in Tables 6.1 and 6.2. In either case, we see that our method outperforms previous approaches by not only increasing overall prediction accuracy but also by increasing consistency. In Figures 6.4 and 6.5, we show illustrative examples of our approach on the

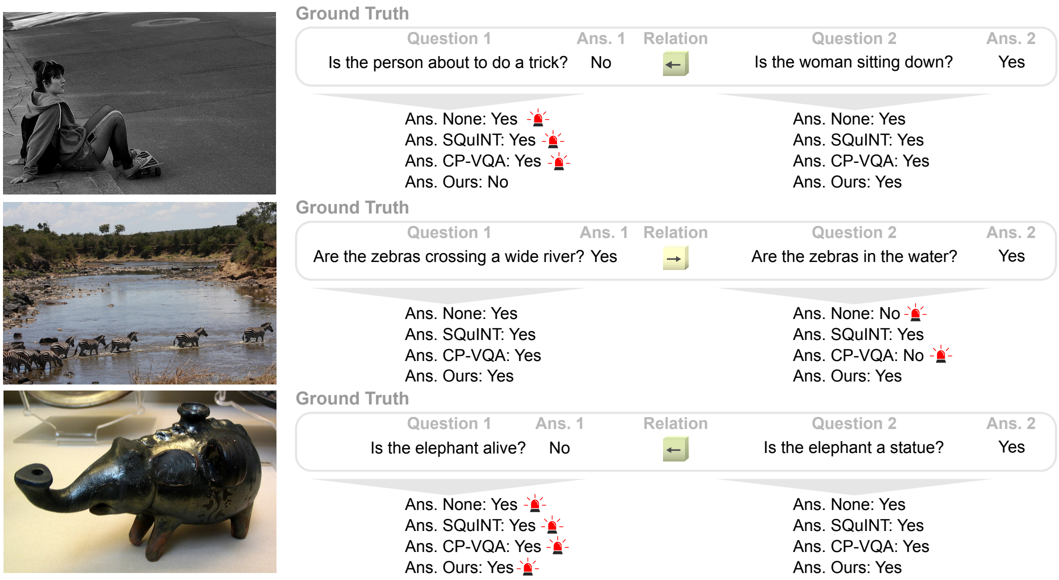


FIGURE 6.4: Qualitative examples from the Introspect dataset using BAN as backbone. Red siren symbols indicate inconsistent cases.

Introspect and DME datasets, respectively (see additional examples in Appendix D).

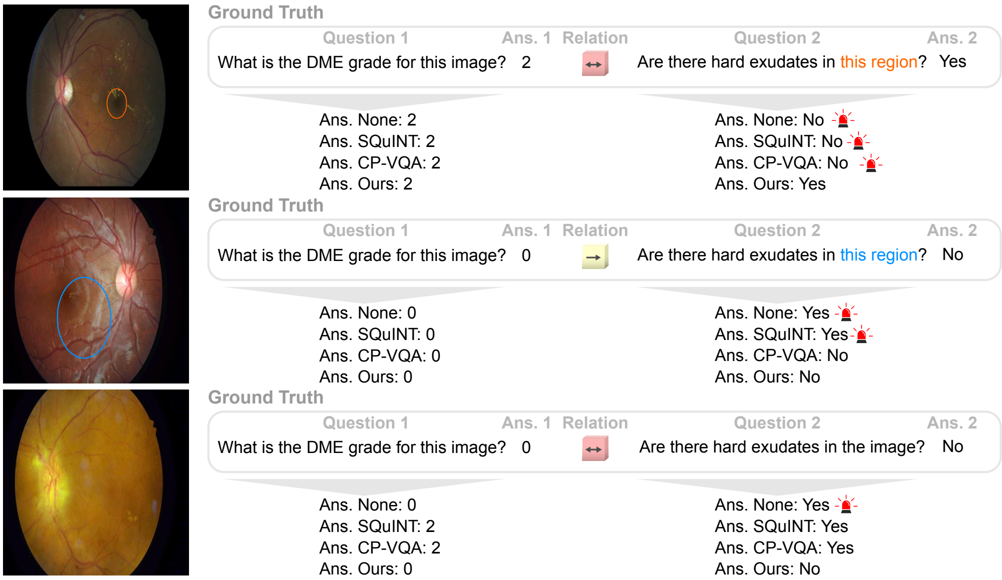
In Table 6.1, we also show the performance of the state-of-the-art LXMERT VQA model when combined with our method. In this case, too, we see that our method provides increased performance via consistency improvements. Here we investigate the performance induced when flipping the answers of one of the members of each related pair at test time. Suppose implication labels are present, either by manual annotation or by LI-MOD. In that case, a trivial manner of correcting an inconsistent QA pair of binary answers is to flip or negate

Model	Cons. Method	Acc.	Cons.
BAN	None	67.14±0.10	69.45±0.17
	SQuINT [40]	67.27±0.19	69.87±0.45
	CP-VQA [149]	67.18±0.24	69.52±0.45
	Ours ($\lambda = 0.01$)	**67.36±0.19**	70.38±0.39
LXMERT	None	75.10±0.10	76.24±0.63
	Random flip	69.67±1.24	75.99±3.91
	Flip first	73.81±0.47	71.94±2.82
	Flip second	65.82±1.03	87.56±2.51
	Ours	**75.17±0.08**	78.75±0.21

TABLE 6.1: Results of different consistency methods on the Introspect dataset using two different VQA models: **Top:** BAN, **bottom:** LXMERT. In the case of LXMERT, we show the impact of randomly flipping the answer of either the first or the second question for related pairs. Similarly, *flip first* and *flip second* refer to flipping the answer to the first and second question in inconsistent pairs, respectively.

Model	Consis. Method	Accuracy					Consistency
		all	grade	whole	macula	region	
MVQA	None	81.15±0.49	78.17±2.07	83.44±1.87	87.25±1.20	80.38±2.02	89.95±3.20
	SQuINT [40]	80.58±0.78	77.48±0.40	82.82±0.74	85.34±0.87	80.02±1.03	89.39±2.12
	CP-VQA [149]	83.49±0.99	**80.69±1.30**	84.96±1.14	87.18±2.18	**83.16±1.09**	94.20±2.15
	Ours ($\lambda = 0.25$)	**83.59±0.69**	80.15±0.95	**86.22±1.67**	**88.18±1.07**	82.62±1.02	95.78±1.19

TABLE 6.2: Comparison of methods on the DME dataset with common MVQA backbone. Accuracy and consistency are reported for all questions, as well as for different medically relevant sub-question categories: grade, whole, macula and region.



Ground Truth

Question 1	Ans. 1	Relation	Question 2	Ans. 2
What is the DME grade for this image?	2	↔	Are there hard exudates in this region?	Yes

Ans. None: 2
Ans. SQuINT: 2
Ans. CP-VQA: 2
Ans. Ours: 2

Ans. None: No 🚨
Ans. SQuINT: No 🚨
Ans. CP-VQA: No 🚨
Ans. Ours: Yes

Ground Truth

Question 1	Ans. 1	Relation	Question 2	Ans. 2
What is the DME grade for this image?	0	→	Are there hard exudates in this region?	No

Ans. None: 0
Ans. SQuINT: 0
Ans. CP-VQA: 0
Ans. Ours: 0

Ans. None: Yes 🚨
Ans. SQuINT: Yes 🚨
Ans. CP-VQA: No
Ans. Ours: No

Ground Truth

Question 1	Ans. 1	Relation	Question 2	Ans. 2
What is the DME grade for this image?	0	↔	Are there hard exudates in the image?	No

Ans. None: 0
Ans. SQuINT: 2
Ans. CP-VQA: 2
Ans. Ours: 0

Ans. None: Yes 🚨
Ans. SQuINT: Yes
Ans. CP-VQA: Yes
Ans. Ours: No

FIGURE 6.5: Examples from the DME dataset and comparison of methods. Red siren symbols indicate inconsistent cases. DME is a disease that is staged into grades (0, 1 or 2), which depend on the number of visual pathological features of the retina. **Top** and **middle:** Although all methods correctly predict the answer to the first question, some inconsistencies appear when a necessary condition is false. **Bottom**: Only the None baseline produces an inconsistency. Note that SQuINT and CP-VQA's answers do not produce inconsistent pairs because both questions were answered incorrectly, and those answers ("2" and "yes") respect all known relations.

one of the answers. This is far simpler than our proposed method as it permits training the VQA model with the standard VQA loss. Having obtained the answers from the model when $\lambda = 0$, we identify the related pairs and then flip the answers (1) either randomly, (2) of the first QA or (3) of the second QA. By including the flipping baselines, we confirm that the added complexity in training our method results in improved accuracy compared to merely correcting inconsistencies post-hoc. Increases in consistency at the expense of accuracy are explained by the fact that an inconsistent QA pair guarantees that one of the two answers is incorrect, but correcting the inconsistency does not necessarily fix the incorrect answer. This phenomenon is particularly noticeable in the flipping baselines, as they fix inconsistencies without considering their correctness.

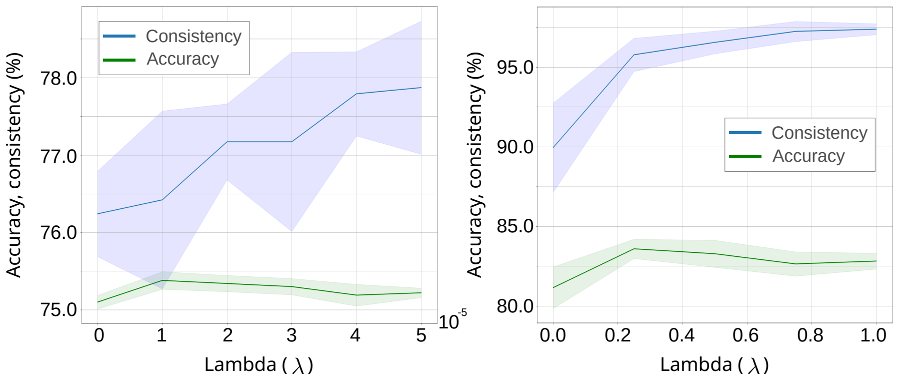


FIGURE 6.6: Behavior of the accuracy and consistency as a function of λ with 95% confidence intervals. **Left:** LXMERT trained on the Introspect dataset (5 models with random seeds for each value of λ). **Right:** MVQA trained on the DME dataset (10 models with random seeds for each λ).

In general, we observe that training LXMERT with our consistency loss provides performance gains. Indeed, while random flipping based on LI-MOD clearly deteriorates the performance of LXMERT, so does flipping the first or second answers. This implies that our proposed method indeed leverages the predictions of LI-MOD to make LXMERT more consistent as it improves both model accuracy and consistency.

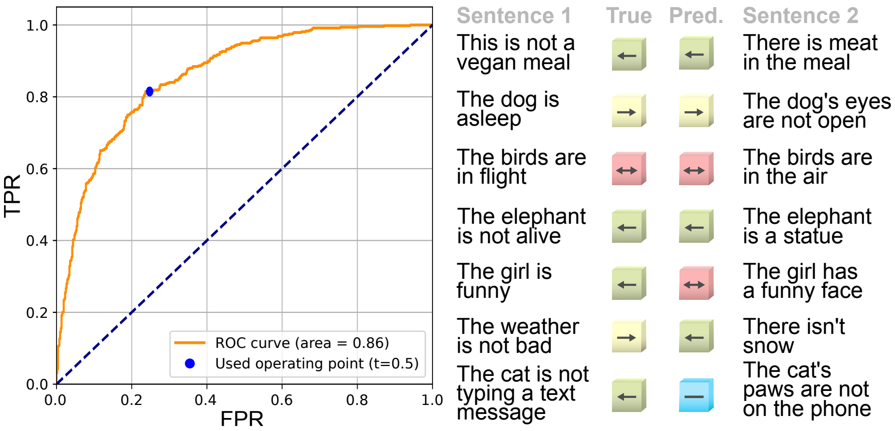
Sensitivity of λ

We now show the sensitivity of our method and its relation to λ. We evaluate the performance of our method for different values of λ to understand the behavior of the performance, both in terms of accuracy and consistency.

Fig. 6.6 shows the accuracy and consistency of LXMERT and MVQA for different values of λ. The difference in the ranges of the values is due to the relative magnitude of the loss function terms and depends on the used loss functions (*e.g.*, binary and non-binary cross-entropy) and the ground-truth answer format (*i.e.*, soft scores for LXMERT, as mentioned in Sec. 6.3.3).

In general, we observe very similar behavior for the accuracy, which increases and then slowly decreases as λ increases. We sustain that the maximum value the accuracy can reach is established by the number of related pairs that are still inconsistent after training with $\lambda = 0$. In other words, the limitations in size impose a limit on how much our method can improve the accuracy. For LXMERT on Introspect, for instance, our model corrected 4,553 (78.9%) of the 5’771 existing inconsistencies and introduced new inconsistencies by mistakenly altering 1,562 (3.5%) of the 44,111 consistent samples.

Regarding consistency, we observe a constant increase as λ increases. The simultaneous decrease in accuracy as λ increases suggests that the relative weight of the consistency loss dominates so that the model no longer focuses on optimizing the cross-entropy. Since it is possible to be consistent without answering correctly, the optimization process results in an



Sentence 1	True	Pred.	Sentence 2
This is not a vegan meal	←	←	There is meat in the meal
The dog is asleep	→	→	The dog's eyes are not open
The birds are in flight	↔	↔	The birds are in the air
The elephant is not alive	←	←	The elephant is a statue
The girl is funny	←	↔	The girl has a funny face
The weather is not bad	→	←	There isn't snow
The cat is not typing a text message	←	—	The cat's paws are not on the phone

FIGURE 6.7: **Left:** Receiver Operating Characteristic (ROC) for the entailment class of our LI-MOD in validation. **Right:** Qualitative examples of LI-MOD's predictions.

increase in consistency at the expense of accuracy for higher values of λ. However, it is clear from these results that there is a set of λ values for which both metrics improve.

LI-MOD Performance

We report that the finetuning of BERT on the subset of annotated relations from Introspect produced 78.67% accuracy in the NLI task. We analyze the performance of this model for entailment and report an AUC value of 0.86, which indicates good generalization capability considering that only $\approx$ 2% of the dataset was annotated with relations. In addition, the overlap in the QA pairs between the train and validation sets of the Introspect dataset is only 1.12% for binary questions. This shows that our LI-MOD is generalizing to variations in questions and to new combinations of QA pairs. Fig. 6.7 shows the ROC curve for entailment and examples of LI-MOD's predictions. Some of the observed sources of errors in LI-MOD include negations, unusual situation descriptions (*e.g.*, a cat typing a text message), and image-specific references (*e.g.*, "is *this* animal real?").

6.4 Conclusion

In this paper, we propose a model-agnostic method to measure and improve consistency in VQA by integrating logical implications between pairs of questions in the training process. We also present a method to infer implications between QA pairs using a transformer-based language model. We conduct experiments to validate the generalizability and robustness of our consistency loss against several baselines and across different datasets. Our results show that our method reduces incoherence in responses and improves performance. Future work includes creating a larger dataset with human-annotated relations to use as a general-purpose relations database for VQA training.

Discussion and Future Work Part III

7 Discussion and Conclusion

The previous four chapters presented solutions addressing two specific challenges in Med-VQA: localized questions and consistency. On the one hand, localized questions (*i.e.*, inquiries about specific regions of an image) enhance the ability to perform localized assessments of image contents, proving particularly valuable in diagnosing and offering second opinions on suspicious regions. Furthermore, having the possibility of asking localized questions has implications for model evaluation, allowing for the examination of agreement both within localized questions and between local and global questions. This approach provides more nuanced insights into the model's actual visual understanding. On the other hand, consistency directly involves the quality of the model's reasoning. Avoiding contradictions and correctly quantifying them becomes a crucial aspect of Med-VQA models, influencing their trustworthiness and potential applicability in the medical practice.

In this chapter, we provide a summary of the findings of the presented works, and discuss their relevance, limitations and significance.

7.1 Discussion

7.1.1 Localized Questions

In Chapters 3 and 4, we introduced two methods that enable asking localized questions in VQA, with a focus on medical images. The first approach (Chapter 3) utilizes the traditional guided-attention mechanism, allowing the model to learn, for a given question, which parts of the image are most relevant. We employ a binary mask to integrate the region's location information in such a way that the model initially considers the evidence from the whole image, and then restricts the attention to the region. In the second approach (Chapter 4), we extend localized questions to MLLMs by creating targeted visual prompts, involving a customized visual prompt with information about the region and its context in the image.

In both of the presented works, the significance of context became apparent in the results, to the extent that a model can rely solely on context to achieve high performance, specially when the regions encompass only a small portion of a much larger object (Sec. 4.3.4). While this observation is evident, it raises the question of how much emphasis should be placed on context when answering a localized question. That is, to what extent should the global understanding of the image influence the final answer to the localized question. We argue that this depends on the size of the region relative to the object(s) about which the question is posed. In cases where the region fully contains the object, more importance could be assigned to the region rather than the context. In other scenarios, our localized attention method (Chapter 3) suggests that the context should be leveraged in a sequential manner, mirroring the way a human would answer the question: first considering the entire image in relation to the question to identify relevant structures, and then focusing attention on the region to make a more detailed decision about its contents. Our method in Chapter 3 replicates this by injecting the binary mask of the target region into the attention mechanism. One limitation of this way of incorporating the region is that the sub-sampling process of the mask can remove details from the contour of the region, potentially causing errors specially when the region is not rectangular. Our second methods mitigates this issue by allowing the model to separately analyze both the entire image and the region.

In Chapter 4, the method presented demonstrates its effectiveness in integrating region-based queries into MLLMs. In essence, the findings align closely with those of our localized attention method. Specifically, the targeted visual prompting method proves to be more effective in datasets like DME-VQA or RIS-VQA, where both the region's contents and the context are mutually crucial for answering the questions. These datasets also exhibit fewer spurious correlations between the region's location/size and its answer, a characteristic that, in the case of INSEGCAT-VQA, leads to certain baselines relying on shortcuts.

Another crucial aspect to consider regarding localized questions is their applicability to real world scenarios in general VQA vs. medical VQA. In this regard, we sustain that applications in the medical domain are easier to envision, due to their potential usefulness in diagnosis, where a localized understanding of the image tends to be more relevant. In the presented works, we make use of three medical datasets. The surgical datasets (RIS-VQA and INSEGCAT-VQA) were created due to the availability of the segmentation annotations, but they do not entirely illustrate a real-world application of localized questions, since the usefulness of the questions is restricted to the detection of surgical instruments.

For natural images, region-defined questions tend to be asked less, in part because of the human tendency to point instead of selecting a region [165]. This, however, does not imply that this type of questions is irrelevant for natural images. As discussed earlier, the evaluation of a model's compositional understanding of and reasoning about the reality expressed by an image is an important emerging field. One possible application of localized questions for natural images is the evaluation of a model's reasoning, as discussed later in Sec. 7.1.3.

7.1.2 Consistency Enhancement

Chapters 5 and 6 introduced two methodologies for enhancing consistency in VQA models. Both approaches leverage a specialized loss function term during training to penalize instances of inconsistency. The primary distinction between the two lies in the definition of consistency and the mathematical function employed to formulate the specialized loss term. In the first method (Chapter 5), questions are categorized as either main (reasoning) or sub (perception), based on the abstract reasoning level required from the model to answer them. This categorization is used to penalize inconsistent cases with a loss term employing cross-entropy as a measure of correctness. In the second method (Chapter 6), a more general framework is proposed. Here, QA pairs are treated as propositions, enabling the application of a more general definition of consistency to pairs linked by an implication relation. Inconsistent cases are then penalized using a loss function that provides high values whenever inconsistent cases occur, contributing to the overall robustness of the model.

Given the pivotal role of consistency in reasoning [40, 180, 188], models exhibiting fewer contradictory or inconsistent answers are perceived as better reasoners. A model that answers “yes" to both “is the pizza vegetarian?" and “is there chicken on the pizza?" is indicative of reasoning issues. As observed, the origin of such contradictions may reside in one or more VQA elements. The language model might struggle with accurate associations, the vision encoder could generate erroneous features, or the block that combines language and vision might encounter difficulties in aligning features. Additionally, the issue might stem from the data itself, where the dataset's distribution affects the model's ability to develop a comprehensive understanding of terms like "vegetarian" due to insufficient examples or biases present in the dataset. In our presented approaches, we adopt a more comprehensive approach, considering the entire VQA model, and strive to enhance consistency for the given dataset.

To enhance consistency, it is crucial to establish an appropriate definition that is both general enough to encompass any pair of QA pairs and robust enough to avoid under and overcounting inconsistent cases. In Chapter 5, following [40], we use a definition of consistency based on the distinction between reasoning and perception questions. Under this definition, two QA pairs are inconsistent for a given image if the answer to the main question is correct while the answer to the sub-question is incorrect. One limitation of this definition is its subjectivity in distinguishing between reasoning and perception; for instance, the question “is the car damaged?" could be perceived as either reasoning or perception, depending on the image depicting the car. A picture of a car with scratches and dents would require simply the detection of such deformations (perception), whereas a picture of a car in an accident scene with oil leaks would require the composition of various perception tasks and prior knowledge.

Another flaw in this definition is its exclusive focus on questions, neglecting the consideration of answers. Furthermore, it assumes an implication relation from the main question q_m to the sub-question q_s (i.e., $q_m \rightarrow q_s$). This assumption is not always valid and stems from the disregard of the answers in the categorization process. For instance, consider the pairs (q_m :“is the person about to do a trick?", a_m : “no") and (q_s :“is the person sitting?", a_s : “yes"). In this scenario, the actual implication relation, considering both questions and answers, is $(q_m, a_m) \leftarrow (q_s, a_s)$. If the model predicts “no" to both questions (*i.e.*, the sub-question is incorrect but the main question is correct), the pair, according to this definition, would be counted as inconsistent. This, however, overlooks the fact that there is no inherent contradiction between a person not being about to do a trick and a person not sitting.

To address these issues, Chapter 6 introduces a general definition of consistency for VQA that relies on the specific modal relation that exists between QA pairs, which are treated as propositions. According to this definition, a pair of QA is considered inconsistent if the propositions implied by them cannot simultaneously be true. This definition is useful for imposing consistency at training time as well as for measuring consistency, as demonstrated in our study. However, a drawback of this approach lies in the assignment of implications to pairs of propositions, as the validity of such implications is not universally applicable in many cases. For example, in the pair, $p =$ (“is it summer?", “no") and $q =$ (“is there snow?", "yes"), the implication $p \leftarrow q$ may generally hold, but exceptions exist (*e.g.*, a glacier in Switzerland or a city in Norway) where the presence of snow does not necessarily negate the occurrence of summer. In this case, we argue that adopting the most general implication ($p \leftarrow q$) is more advantageous for the model. The model would naturally learn exceptions to these general rules based on the samples within the dataset. Consequently, the accuracy and quality of data annotations become critical for developing models with enhanced consistency.

When examining consistency, a critical aspect that deserves analysis is its relationship with accuracy. Since a higher consistency does not necessarily imply increased accuracy, it is worth breaking down the potential interactions that can occur for a pair of QA (or propositions). Considering the related binary QA pairs $(q_1, a_1) \rightarrow (q_2, a_2)$, as described in Chapter 6, the pair will be deemed inconsistent when the model produces answers $\hat{a}_1$ and $\hat{a}_2$ satisfying $\hat{a}_1 = a_1$ and $\hat{a}_2 = \neg a_2$. In the ideal scenario, the model would correct the answer to q_2 from $\neg a_2$ to a_2 while maintaining $\hat{a}_1 = a_1$, thereby resolving the inconsistency and improving accuracy simultaneously. However, two other outcomes could also render the pair consistent: $(\neg a_1, a_2)$

and $(\neg a_1, \neg a_2)$. The former enhances consistency while maintaining the same accuracy, and the latter improves consistency at the expense of accuracy. In contrast to previously proposed methods that compromise accuracy [40, 170, 175], our proposed methods demonstrate that the model is encouraged to rectify inconsistencies by altering the incorrect answer in the pair, as opposed to changing both answers or modifying the one corresponding to the sufficient condition. This is a relevant outcome under the consideration that regularizers can reach convergence after the main loss term [179].

Regarding the mathematical definitions of the proposed terms, certain limitations can be identified. In the method presented in Chapter 5, the mathematical function requires a hyperparameter γ to determine at which value of the cross-entropy of the main question the penalty should be disabled. While effective in enhancing consistency, this loss term requires a certain level of heuristic exploration to find the most effective value for γ, making the process somewhat tedious. The generality of our second method removes the need to find the optimal value of a hyperparameter other than the loss term gain. This, coupled with the absence of abrupt changes in the function, constitutes a significant improvement.

7.1.3 Bridging Locality and Consistency

Localized questions and consistency, as discussed, are individually significant for localized examination of images and reasoning, respectively. We will now explore how incorporating localized questions into consistency enhancement can broaden the possibilities for evaluating reasoning and visual understanding.

The primary advantage of this combination of localized questions and consistency is the expansion of consistency evaluation from exclusively global-to-global to global-to-local and local-to-local. Here *global* refers to questions about the entire image, and *local* refers to localized questions. With this expansion, the consistency evaluation becomes more detailed and localized, contributing to the reliability of the models. Consider, for instance, the questions from the DME-VQA [149] that refer to the presence of hard exudates in the entire image and in a specific region. Here, once again, we emphasize the dependence of consistency on the answers since answering “yes" to the global question admits both “yes" and “no" responses to the local question, depending on the location of the region. However, answering “no" to the global question imposes, from the standpoint of consistency, a constraint in the answer to the local questions. In this scenario, and assuming the prediction for the global question is correct, the model should answer “no" to all region-based questions that inquire about the same biomarker. We contend that this type of compositional evaluation can make the consistency assessment of a model more robust, as its understanding of the image is tested at both global and local levels, revealing possible shortcuts or perception errors.

In the local-to-local context, certain relevant reasoning failures could be detected (or ensured to be absent). Revisiting the questions about hard exudates in the DME-VQA dataset, we underscore the importance of answers to identical questions concerning cases where a region

r_1 contains another one, r_2, with $area(r_1) > area(r_2)$. Given the question q = “are there hard exudates in this region?" we observe that an affirmative answer to q_{r_2} implies an affirmative answer to q_{r_1} and that a negative answer to q_{r_1} implies a negative answer to q_{r_2}. These implications enable the identification of inconsistencies in which the model alters its prediction when exposed to more or less context, respectively. This, however, should be analyzed carefully since the mentioned implications do not apply to all types of images/questions.

In both extensions of consistency mentioned, namely global-to-local and local-to-local, the approaches presented in this work emphasize the need for models capable of answering localized questions by considering context while also respecting the precise boundaries and contents of regions. Simultaneously, these models should assimilate knowledge about implication violations to enhance performance and overall reasoning capabilities.

7.2 Conclusion

In conclusion, this thesis has delved into the principles of VQA in the medical domain, focusing on two key aspects: enabling localized queries and enhancing consistency. The exploration of these dimensions of Med-VQA contributes to a nuanced understanding of medical image interpretation and reasoning. The devised methodologies, from introducing localized attention mechanisms and a visual prompting technique for MLLMs to proposing a logic-centric method for consistency, mark noteworthy advances in addressing the challenges posed by VQA in medical scenarios. Furthermore, several Med-VQA datasets were created and made publicly available, fostering research efforts within the research community in the domains of localized questions and consistency.

The work on localized questions introduces novel ways for users to query specific regions of medical images, supporting a more targeted and informative interaction with the model. This has implications for clinical applications, offering potential benefits in diagnosis, second opinions, and overall medical image analysis.

Conversely, the in-depth considerations regarding consistency in VQA models reveal its critical role in improving reasoning capabilities. The presented methods penalize inconsistent cases at training time, resulting in higher consistency and overall performance during inference, thereby showing the adequacy of the enhancement techniques.

By intertwining the realms of localized queries and consistency, this work paves the way for a comprehensive approach to robustness in Med-VQA. This combination not only refines the assessment but also exposes the models to diverse challenges that go beyond global information, ensuring a more robust and trustworthy performance.

In essence, this thesis contributes to the evolving landscape of medical VQA by introducing approaches that enhance interpretability, reasoning, and reliability. As the field continues to advance with the adoption of larger models, the insights and methodologies presented herein offer valuable perspectives, emphasizing the ongoing quest for more accurate, consistent, and context-aware Med-VQA systems.

8 Future Work

The evolution of VQA and Med-VQA has been remarkably rapid in the last years, both at the architectural and data levels. Initially, the structure comprised a simple combination of an RNN, a CNN, a multiplication operation, and a classifier. Over time, this has transformed into sophisticated MLLMs with billions of parameters, forming complex stacks of layers with attention mechanisms at their core.

On the data front, datasets have expanded to encompass millions of questions, addressing existing biases and incorporating additional information such as scene graphs [29]. The LLMs responsible for reasoning in MLLMs are trained on internet-scale datasets with hundreds of billions of words. Despite these advancements, visual understanding and multimodal reasoning remain pertinent topics [100, 215–217] that deserve the attention of the research community.

Taking into account these advancements and the works presented in this thesis, this chapter delineates directions and considerations for future work in the domains of consistency and localized questions.

8.1 Localized Questions

Regarding the data used to train models that answer localized questions, a natural progression from our current focus on binary questions is the integration of non-binary inquiries. Our study was limited to binary questions due to the absence of publicly available VQA datasets featuring questions about regions or other datasets with spatial annotations beyond segmentation masks. In future developments, the formulation of questions could include a spectrum of topics, including anatomical descriptions (*e.g.*, “which organ is in this region?"), comparisons (*e.g.*, “how does this region compare to a healthy counterpart in the image?"), spatial relationships (*e.g.*, “what structures or organs are adjacent to this region?"), functional questions (*e.g.*, “what is the organ in the region responsible for?"), pathological descriptions (*e.g.*, “what pathological findings or abnormalities are in this region?"), diagnostic questions (*e.g.*, “what diagnostic information can be derived from the organ in this region?"), etc.

Creating datasets for such diverse questions would require active involvement from clinical experts in the generation of each QA pair, as well as carefully considering inter-expert variability to minimize biases in the datasets. Alternatively, leveraging heavily annotated medical datasets could automate the generation of questions. Ideally, a wide range of modalities such as CT, MRI, X-Ray, and OCT, among others, should be included to enhance the applicability of Med-VQA models in clinical practice. Moreover, questions and answers should accurately reflect the style and specialized terminology employed by medical professionals, a goal that could be feasibly attained through the utilization of LLMs.

Prioritizing the minimization of biases in the data is imperative, particularly considering the potential impact of object size on answer distribution. For instance, in an image with few and small lesions, generating localized questions about healthy tissue might be easier than those about abnormal tissue. These data balance considerations are crucial to reduce model reliance on spurious correlations based on object location or size.

Regarding our method from Chapter 3 for localized questions, further developments could involve exploring alternative architectures in the text and image embedding blocks. For instance, for the text encoder, investigating the efficacy of the RWKV architecture [218] combining RNNs with transformers could be a possibility. Additionally, investigating alternative multimodal fusion approaches, like bilinear pooling, could potentially improve performance.

A more drastic modification could involve adopting an attention pyramid, where attention maps are created at different depths of the visual encoder's features. This might address the potential loss of region detail when applying the resized binary mask to visually attended features when the region shape is complex.

Furthermore, delving deeper into the role of glimpses for localized questions could refine the model's handling of redundancy. Observations during experimental development showed instances where one glimpse focuses on the object mentioned in the question while another highlights a different part of the image. Investigating this aspect can contribute to a more nuanced and effective localized questioning approach.

Concerning our efforts in Chapter 4, alternative approaches might explore the use of CNNs for encoding the image, either independently or in conjunction with ViTs. Such an approach could facilitate a direct mapping of visual tokens to the input image and potentially allow for a combination of both proposed methods. In this scenario, visual features could undergo filtering based on the target regions (localized attention), enabling the LLM to receive locally attended visual tokens.

8.2 Consistency Enhancement

Similar to the scenario of localized questions, the incorporation of non-binary questions for consistency enhancement represents a crucial advancement. This extension allows for a more comprehensive examination of the model's reasoning capabilities and language understanding. In our exploration using the Introspect-VQA dataset (Chapter 6), we exclusively employed binary questions due to the intricacies associated with assigning modal relations to non-binary QA pairs. However, the annotations regarding implications could be acquired during dataset creation or generated by a sophisticated LLM. Leveraging the recent progress in LLMs, these models could serve as auxiliary tools for logical reasoning/knowledge. By learning from extensive text data and building associations, LLMs more closely resemble human learning of implications, which takes place in different forms (experience, deductive and inductive reasoning, error correction, etc.). This, linked to the work of this thesis, corresponds to upgrading LI-MOD in the method from Chapter 6. An added advantage of using a pre-trained LLM is its potential for zero-shot operation, drawing on the knowledge abstracted from extensive training data.

For the work presented in Chapter 5, envisioning a dynamic tuning of the hyperparameter γ could be explored. However, this should be performed under the revised definition of consistency (Chapter 6), considering the limitations associated with the main-sub categorization, as discussed earlier. Finding new functions for the loss term could also represent an interesting avenue for further development.

Refining the consistency definitions offers another potential direction, with a focus on considering the context provided in the image to weigh the applied implications. That is, a more robust definition could relax the implication relation based on the contents of the image. For instance, in the previously examined example (“is it summer?", “no") ← (“is there snow?", "yes"), the implication relation holds for most cases, but exceptions exist. A more nuanced definition of consistency for VQA could adapt the implication relation based on the specifics of the image contents: if the image shows snow at the top of a mountain, the chances that the relation does not hold increase as compared to an image showing snow on a street in New York.

With respect to the method in Chapter 6, testing more advanced architectures, such as

OFA [120], BEIT-3 [219], or other VLMs, could be pursued to evaluate the method's effectiveness. In this context, the increased capacity and pre-training of these models are anticipated to lead to enlarged consistency. Our method is expected to provide additional benefits in addressing contradictions beyond what increased overall performance can bring.

Finally, exploring human-in-the-loop approaches represents another avenue where models could learn to avoid contradictions by leveraging error feedback provided by humans.

8.3 Bridging Locality and Consistency

We now explore potential avenues for future research at the intersection of localized questions and consistency enhancement.

In the realm of real-world clinical applications, a promising area for investigation involves evaluating the trustworthiness of VQA models. This could entail engaging medical experts with VQA models featuring varying levels of consistency to assess their trust in the system. Questions about regions would be included, allowing the experts to probe the model in a more dynamic and interactive way. Such interaction would provide insights into the types of contradictions that may impact specialists' reluctance to adopt these models in clinical practice.

Another avenue pertinent to clinical environments is the development of MLLM-based medical assistants with consistency enhancement mechanisms. Due to the availability of large datasets, this could be conceived first for pathology and radiology images, building on recent efforts [220–222], but should then be expanded to other modalities and should integrate the option to ask questions about user-defined regions. These assistants would enable users to pose questions of any kind (text-only, image, or region) while ensuring coherence in responses.

With reference to dataset creation, there is a need for the development of more robust datasets incorporating both localized questions and relation annotations. This development would address the challenges outlined previously for each scenario, ensuring the availability of diverse QA pairs that put the model to the test in terms of relations between propositions that also involve prior knowledge (*e.g.*, presence of biomarkers that imply a certain disease).

One significant advancement over implication relations involves the inclusion of more than two propositions. Scenarios could be envisioned where multiple propositions collectively lead to a conclusion (*i.e.*, $A_1 \wedge A_2 \wedge ... \wedge A_N \rightarrow B$). Here, the propositions A_i, $i = 1, ..., N$ can be treated as one single proposition, allowing the application of the same definition of consistency from Chapter 6. If a model assigns *true* to $A_i, i = 1, ..., N$ but *false* to B, the set of propositions will be considered inconsistent. Importantly, the individual pairs $\{(A_1, B), ..., (A_N, B)\}$ are not considered inconsistent, since the implication relation requires the evaluation of all propositions simultaneously. This can be useful for diagnosis decisions requiring the presence of N biomarkers in the image, or in certain regions. An extension could also be made for cases where the question contains prior information about the patient not present in the image, such as findings from previous images (*i.e.*, longitudinal information) or blood work results. In such scenarios, the question would assert a proposition, and could be included in the consistency evaluation as a complement to the propositions implied by QA pairs.

Enhancing the evaluation of consistency in the context of localized questions may require optimizing metrics. Assigning different weights to inconsistencies based on the extent of the visual information associated with the violated implication relation could be a strategic approach. Deeming local-to-local inconsistencies as more critical than global-to-global inconsistencies seems intuitive, given that the former involve less visual information and should be more manageable for the model to address. Global questions often require higher-level reasoning and the composition of various perception tasks, hence, imposing more significant penalties on seemingly straightforward cases appears justifiable. This, of course, rests on the assumption that the model can answer both types of questions at a comparable level.

Another possible direction for future work in localized questions and consistency is the development of explainability and interpretability methods. These methods aim to summarize the model's performance and attempt to explain its predictions. For instance, global predictions could be deconstructed into local predictions about non-overlapping regions, showcasing the extent to which the model's interpretation of the entire image can be decomposed into an understanding of its constituent parts (compositional VQA). Other approaches could be devised to test a model's comprehension of an image by identifying regions for which the model's answers contradict the response to a global question. Such an approach may uncover objects or structures that the model confuses with those mentioned in the question, contributing to a deeper understanding of model behavior.

Bibliography

[1] Graham Flegg et al. *Numbers through the ages*. Ed. by Graham Flegg. Springer, 1989.

[2] Mike G Edmunds. *The Antikythera mechanism and the mechanical universe*. Vol. 55. 4. Taylor & Francis, 2014, pp. 263–285.

[3] John D North. “The astrolabe”. In: *Scientific American* 230.1 (1974), pp. 96–107.

[4] Cliff Stoll. “When slide rules ruled”. In: *Scientific American* 294.5 (2006), pp. 80–87.

[5] Eugene Eric Kim and Betty Alexandra Toole. “Ada and the first computer”. In: *Scientific American* 280.5 (1999), pp. 76–81.

[6] Alan M Turing. *Computing machinery and intelligence*. Springer, 2009.

[7] Robert M French. “Subcognition and the limits of the Turing test”. In: *Mind* 99.393 (1990), pp. 53–65.

[8] Arthur L Samuel. “Some studies in machine learning using the game of checkers”. In: *IBM Journal of research and development* 3.3 (1959), pp. 210–229.

[9] Frank Rosenblatt. “The perceptron: A probabilistic model for information storage and organization in the brain”. In: *Psychological Review* 65.6 (1958), pp. 386–408. DOI: 10.1037/h0042519.

[10] Seppo Linnainmaa. “The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors”. PhD thesis. Master’s Thesis (in Finnish), Univ. Helsinki, 1970.

[11] Paul J Werbos. “Applications of advances in nonlinear sensitivity analysis”. In: *System Modeling and Optimization: Proceedings of the 10th IFIP Conference New York City, USA, August 31–September 4, 1981*. Springer. 2005, pp. 762–770.

[12] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. “Learning Internal Representations by Error Propagation, Parallel Distributed Processing, Explorations in the Microstructure of Cognition”. In: *Biometrika* 71 (1986), pp. 599–607.

[13] Feng-hsiung Hsu. “IBM’s deep blue chess grandmaster chips”. In: *IEEE micro* 19.2 (1999), pp. 70–81.

[14] Richard Szeliski. *Computer vision: algorithms and applications*. Springer Nature, 2022.

[15] Kunihiko Fukushima. “Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position”. In: *Biological cybernetics* 36.4 (1980), pp. 193–202.

[16] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “Imagenet classification with deep convolutional neural networks”. In: *Advances in neural information processing systems* 25 (2012).

[18] Ashwin Bhandare, Maithili Bhide, Pranav Gokhale, and Rohan Chandavarkar. “Applications of convolutional neural networks”. In: *International Journal of Computer Science and Information Technologies* 7.5 (2016), pp. 2206–2215.

[19] Alexey Dosovitskiy et al. “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *arXiv preprint arXiv:2010.11929* (2020).

[20] Prashant Johri, Sunil K Khatri, Ahmad T Al-Taani, Munish Sabharwal, Shakhzod Suvanov, and Avneesh Kumar. “Natural language processing: History, evolution, application, and future work”. In: *Proceedings of 3rd International Conference on Computing Informatics and Networks: ICCIN 2020.* Springer. 2021, pp. 365–375.

[21] W John Hutchins. “Retrospect and prospect in computer-based translation”. In: *Proceedings of Machine Translation Summit VII.* 1999, pp. 30–36.

[22] Noam Chomsky. “Chomsky”. In: *DANCY, J, A Companion to the Philosophy of* (1896).

[23] Joseph Weizenbaum. “ELIZA—a computer program for the study of natural language communication between man and machine”. In: *Communications of the ACM* 9.1 (1966), pp. 36–45.

[24] Terry Winograd. “What does it mean to understand language?” In: *Cognitive science* 4.3 (1980), pp. 209–241.

[25] William A Woods. “Progress in natural language understanding: an application to lunar geology”. In: *Proceedings of the June 4-8, 1973, national computer conference and exposition.* 1973, pp. 441–450.

[26] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. “A neural probabilistic language model”. In: *Advances in neural information processing systems* 13 (2000).

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).

[28] Soumen Pal, Manojit Bhattacharya, Md Aminul Islam, and Chiranjib Chakraborty. “ChatGPT or LLM in next-generation drug discovery and development: pharmaceutical and biotechnology companies can make use of the artificial intelligence-based device for a faster way of drug discovery and development”. In: *International Journal of Surgery* 109.12 (2023), pp. 4382–4384.

[29] Drew A Hudson and Christopher D Manning. “Gqa: a new dataset for compositional question answering over real-world images”. In: *arXiv preprint arXiv:1902.09506* 3.8 (2019).

[30] Boris Katz, Jimmy Lin, Chris Stauffer, and W Eric L Grimson. Book chapter: “Answering Questions about Moving Objects in Surveillance Videos.” In: *New directions in question answering*. Ed. by Mark T. Maybury. 2003, pp. 145–152.

[31] Hui Yang, Lekha Chaisorn, Yunlong Zhao, Shi-Yong Neo, and Tat-Seng Chua. “VideoQA: question answering on news video”. In: *Proceedings of the eleventh ACM international conference on Multimedia*. 2003, pp. 632–641.

[32] Tom Yeh, John J Lee, and Trevor Darrell. “Photo-based question answering”. In: *Proceedings of the 16th ACM international conference on Multimedia*. 2008, pp. 389–398.

[33] Jeffrey P Bigham et al. “Vizwiz: nearly real-time answers to visual questions”. In: *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. 2010, pp. 333–342.

[34] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. “Vizwiz grand challenge: Answering visual questions from blind people”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 3608–3617.

[35] Danna Gurari, Qing Li, Chi Lin, Yinan Zhao, Anhong Guo, Abigale Stangl, and Jeffrey P Bigham. “Vizwiz-priv: A dataset for recognizing the presence and purpose of private visual information in images taken by blind people”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 939–948.

[36] Kewei Tu, Meng Meng, Mun Wai Lee, Tae Eun Choe, and Song-Chun Zhu. “Joint video and text parsing for understanding events and answering queries”. In: *IEEE MultiMedia* 21.2 (2014), pp. 42–70.

[37] Mateusz Malinowski and Mario Fritz. “A multi-world approach to question answering about real-world scenes based on uncertain input”. In: *Advances in neural information processing systems* 27 (2014).

[38] Donald Geman, Stuart Geman, Neil Hallonquist, and Laurent Younes. “Visual turing test for computer vision systems”. In: *Proceedings of the National Academy of Sciences* 112.12 (2015), pp. 3618–3623.

[39] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. “Vqa: Visual question answering”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 2425–2433.

[40] Ramprasaath R Selvaraju, Purva Tendulkar, Devi Parikh, Eric Horvitz, Marco Tulio Ribeiro, Besmira Nushi, and Ece Kamar. “SQuINTing at VQA Models: Introspecting VQA Models With Sub-Questions”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 10003–10011.

[41] Corentin Kervadec. “Bias and reasoning in visual question answering”. PhD thesis. Université de Lyon, 2021.

[42] Eda Kavlakoglu. *NLP vs. NLU vs. NLG: the differences between three natural language processing concepts - IBM Blog — ibm.com.* https://www.ibm.com/blog/nlp-vs-nlu-vs-nlg-the-differences-between-three-natural-language-processing-concepts/. [Accessed 25-01-2024].

[43] Laurenz Wuttke. *NLP vs. NLU vs. NLG: Unterschiede, Funktionen und Beispiele — datasolut.com.* https://datasolut.com/natural-language-processing-vs-nlu-vs-nlg-unterschiede-funktionen-und-beispiele/. [Accessed 25-01-2024].

[44] Delip Rao and Brian McMahan. *Natural language processing with PyTorch: build intelligent language applications using deep learning.* " O’Reilly Media, Inc.", 2019.

[45] Josef Sivic and Andrew Zisserman. “Efficient visual search of videos cast as text retrieval”. In: *IEEE transactions on pattern analysis and machine intelligence* 31.4 (2008), pp. 591–606.

[46] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. “Efficient estimation of word representations in vector space”. In: *arXiv preprint arXiv:1301.3781* (2013).

[47] Jeffrey Pennington, Richard Socher, and Christopher D Manning. “Glove: Global vectors for word representation”. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP).* 2014, pp. 1532–1543.

[48] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).

[49] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning.* MIT press, 2016.

[50] *What are Recurrent Neural Networks? | IBM — ibm.com.* https://www.ibm.com/topics/recurrent-neural-networks. [Accessed 26-01-2024].

[51] Mike Schuster and Kuldip K Paliwal. “Bidirectional recurrent neural networks”. In: *IEEE transactions on Signal Processing* 45.11 (1997), pp. 2673–2681.

[52] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory”. In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[53] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. “Learning phrase representations using RNN encoder-decoder for statistical machine translation”. In: *arXiv preprint arXiv:1406.1078* (2014).

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

[55] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. “Layer normalization”. In: *arXiv preprint arXiv:1607.06450* (2016).

[56] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate". In: *arXiv preprint arXiv:1409.0473* (2014).

[57] Jay Alammar. *The Illustrated Transformer — jalammar.github.io*. http://jalammar.github.io/illustrated-transformer/. [Accessed 26-01-2024].

[58] Tom Brown et al. "Language models are few-shot learners". In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[59] *What are Large Language Models? | NVIDIA Glossary — nvidia.com*. https://www.nvidia.com/en-us/glossary/large-language-models/. [Accessed 27-01-2024].

[60] Shaden Smith et al. "Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model". In: *arXiv preprint arXiv:2201.11990* (2022).

[61] Josh Achiam et al. "Gpt-4 technical report". In: *arXiv preprint arXiv:2303.08774* (2023).

[62] *Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, the World's Largest and Most Powerful Generative Language Model | NVIDIA Technical Blog — developer.nvidia.com*. https://developer.nvidia.com/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model/. [Accessed 27-01-2024].

[63] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. "Exploring the limits of transfer learning with a unified text-to-text transformer". In: *The Journal of Machine Learning Research* 21.1 (2020), pp. 5485–5551.

[64] Dave Van Veen et al. "Clinical text summarization: Adapting large language models can outperform human experts". In: *arXiv preprint arXiv:2309.07430* (2023).

[65] Marc Cicero Schubert, Wolfgang Wick, and Varun Venkataramani. "Large Language Model-Driven Evaluation of Medical Records Using MedCheckLLM". In: *medRxiv* (2023), pp. 2023–11.

[66] Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, and Anima Anandkumar. "Eureka: Human-Level Reward Design via Coding Large Language Models". In: *arXiv preprint arXiv: Arxiv-2310.12931* (2023).

[67] Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. "R2gengpt: Radiology report generation with frozen llms". In: *Meta-Radiology* 1.3 (2023), p. 100033.

[68] Julian Varas et al. "Innovations in surgical training: exploring the role of artificial intelligence and large language models (LLM)". In: *Revista do Colégio Brasileiro de Cirurgiões* 50 (2023), e20233605.

[69] Devi Prasad Mohapatra, Friji Meethale Thiruvoth, Satyaswarup Tripathy, Sheeja Rajan, Madhubari Vathulya, Palukuri Lakshmi, Veena K Singh, and Ansar Ul Haq. "Leveraging Large Language Models (LLM) for the Plastic Surgery Resident Training: Do They Have a Role?" In: *Indian Journal of Plastic Surgery* 56.05 (2023), pp. 413–420.

[70] Bertalan Meskó. “The impact of multimodal large language models on health care’s future”. In: *Journal of Medical Internet Research* 25 (2023), e52865.

[71] Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and S Yu Philip. “Multimodal large language models: A survey”. In: *2023 IEEE International Conference on Big Data (BigData)*. IEEE. 2023, pp. 2247–2256.

[72] Can Cui et al. “A survey on multimodal large language models for autonomous driving”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2024, pp. 958–979.

[73] BigScience Workshop et al. “Bloom: A 176b-parameter open-access multilingual language model”. In: *arXiv preprint arXiv:2211.05100* (2022).

[74] Kunal Agarwal. *Council Post: Why Optimizing Cost Is Crucial To AI/ML Success — forbes.com*. https://www.forbes.com/sites/forbestechcouncil/2023/09/13/why-optimizing-cost-is-crucial-to-aiml-success/. [Accessed 02-02-2024].

[75] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. *Parameter-Efficient Transfer Learning for NLP*. 2019. arXiv: 1902.00751 [cs.LG].

[76] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. “Lora: Low-rank adaptation of large language models”. In: *arXiv preprint arXiv:2106.09685* (2021).

[77] Xiang Lisa Li and Percy Liang. *Prefix-Tuning: Optimizing Continuous Prompts for Generation*. 2021. arXiv: 2101.00190 [cs.CL].

[78] Shiv Ram Dubey, Satish Kumar Singh, and Bidyut Baran Chaudhuri. “Activation functions in deep learning: A comprehensive survey and benchmark”. In: *Neurocomputing* (2022).

[79] Hossein Gholamalinezhad and Hossein Khosravi. “Pooling methods in deep neural networks, a review”. In: *arXiv preprint arXiv:2009.07485* (2020).

[80] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows*. 2021. arXiv: 2103.14030 [cs.CV].

[81] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. *Training data-efficient image transformers & distillation through attention*. 2021. arXiv: 2012.12877 [cs.CV].

[82] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. “Stacked cross attention for image-text matching”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 201–216.

[83] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. “Referitgame: Referring to objects in photographs of natural scenes”. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 787–798.

[84] Andrej Karpathy and Li Fei-Fei. *Deep Visual-Semantic Alignments for Generating Image Descriptions*. 2015. arXiv: 1412.2306 [cs.CV].

[85] Liam Hebert, Gaurav Sahu, Yuxuan Guo, Nanda Kishore Sreenivas, Lukasz Golab, and Robin Cohen. *Multi-Modal Discussion Transformer: Integrating Text, Images and Graph Transformers to Detect Hate Speech on Social Media*. 2024. arXiv: 2307.09312 [cs.CL].

[86] Haopeng Li, Qiuhong Ke, Mingming Gong, and Tom Drummond. "Progressive Video Summarization via Multimodal Self-Supervised Learning". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Jan. 2023, pp. 5584–5593.

[87] Alec Radford et al. *Learning Transferable Visual Models From Natural Language Supervision*. 2021. arXiv: 2103.00020 [cs.CV].

[88] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. *InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning*. 2023. arXiv: 2305.06500 [cs.CV].

[89] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. *Visual Instruction Tuning*. 2023. arXiv: 2304.08485 [cs.CV].

[90] *GPT-4V(ision) System Card*. https://cdn.openai.com/papers/GPTV_System_Card.pdf. [Accessed 31-01-2024].

[91] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, et al. *Gemini: A Family of Highly Capable Multimodal Models*. 2023. arXiv: 2312.11805 [cs.CL].

[92] Duzhen Zhang, Yahan Yu, Chenxing Li, Jiahua Dong, Dan Su, Chenhui Chu, and Dong Yu. *MM-LLMs: Recent Advances in MultiModal Large Language Models*. 2024. arXiv: 2401.13601 [cs.CL].

[93] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. *A Survey on Multimodal Large Language Models*. 2023. arXiv: 2306.13549 [cs.CV].

[94] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. *BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models*. 2023. arXiv: 2301.12597 [cs.CV].

[95] Wenhai Wang et al. *VisionLLM: Large Language Model is also an Open-Ended Decoder for Vision-Centric Tasks*. 2023. arXiv: 2305.11175 [cs.CV].

[96] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. "Otter: A multi-modal model with in-context instruction tuning". In: *arXiv preprint arXiv:2305.03726* (2023).

[97] Peng Gao et al. *LLaMA-Adapter V2: Parameter-Efficient Visual Instruction Model*. 2023. arXiv: 2304.15010 [cs.CV].

[98] Renrui Zhang et al. *LLaMA-Adapter: Efficient Fine-tuning of Language Models with Zero-init Attention*. 2023. arXiv: 2303.16199 [cs.CV].

[99] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. *VideoChat: Chat-Centric Video Understanding.* 2024. arXiv: 2305.06355 [cs.CV].

[100] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. "Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs". In: *arXiv preprint arXiv:2401.06209* (2024).

[101] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. "Finetuned language models are zero-shot learners". In: *arXiv preprint arXiv:2109.01652* (2021).

[102] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. "A survey for in-context learning". In: *arXiv preprint arXiv:2301.00234* (2022).

[103] Feifan Liu, Yalei Peng, and Max P Rosen. "An effective deep transfer learning and information fusion framework for medical visual question answering". In: *International Conference of the Cross-Language Evaluation Forum for European Languages.* Springer. 2019, pp. 238–247.

[104] Zhibin Liao, Qi Wu, Chunhua Shen, Anton Van Den Hengel, and Johan Verjans. "Aiml at vqa-med 2020: Knowledge inference via a skeleton-based sentence mapping approach for medical domain visual question answering". In: (2020).

[105] Deepak Gupta, Swati Suman, and Asif Ekbal. "Hierarchical deep multi-modal network for medical visual question answering". In: *Expert Systems with Applications* 164 (2021), p. 113993.

[106] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. "Multimodal compact bilinear pooling for visual question answering and visual grounding". In: *arXiv preprint arXiv:1606.01847* (2016).

[107] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. "Bilinear attention networks". In: *Advances in neural information processing systems* 31 (2018).

[108] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering". In: *Proceedings of the IEEE international conference on computer vision.* 2017, pp. 1821–1830.

[109] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. "Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering". In: *IEEE transactions on neural networks and learning systems* 29.12 (2018), pp. 5947–5959.

[110] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. "Mutan: Multimodal tucker fusion for visual question answering". In: *Proceedings of the IEEE international conference on computer vision.* 2017, pp. 2612–2620.

[111] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. *Stacked Attention Networks for Image Question Answering.* 2016. arXiv: 1511.02274 [cs.LG].

[112] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. "Bottom-up and top-down attention for image captioning and visual question answering". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 6077–6086.

[113] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. "Faster r-cnn: Towards real-time object detection with region proposal networks". In: *Advances in neural information processing systems* 28 (2015).

[114] Huijuan Xu and Kate Saenko. "Ask, attend and answer: Exploring question-guided spatial attention for visual question answering". In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*. Springer. 2016, pp. 451–466.

[115] Hao Tan and Mohit Bansal. "Lxmert: Learning cross-modality encoder representations from transformers". In: *arXiv preprint arXiv:1908.07490* (2019).

[116] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. "In defense of grid features for visual question answering". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 10267–10276.

[117] Zujie Liang, Weitao Jiang, Haifeng Hu, and Jiaying Zhu. "Learning to contrast the counterfactual samples for robust visual question answering". In: *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*. 2020, pp. 3285–3292.

[118] Damien Teney, Ehsan Abbasnejad, and Anton van den Hengel. "Unshuffling data for improved generalization". In: *arXiv preprint arXiv:2002.11894* (2020).

[119] Aisha Urooj Khan, Hilde Kuehne, Kevin Duarte, Chuang Gan, Niels Lobo, and Mubarak Shah. *Found a Reason for me? Weakly-supervised Grounded Visual Question Answering using Capsules*. 2021. arXiv: 2105.04836 `[cs.CV]`.

[120] Peng Wang et al. *OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework*. 2022. arXiv: 2202.03052 `[cs.CV]`.

[121] Jean-Baptiste Alayrac et al. *Flamingo: a Visual Language Model for Few-Shot Learning*. 2022. arXiv: 2204.14198 `[cs.CV]`.

[122] Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven C. H. Hoi. *From Images to Textual Prompts: Zero-shot VQA with Frozen Large Language Models*. 2023. arXiv: 2212.10846 `[cs.CV]`.

[123] Haibi Wang and Weifeng Ge. *Q&A Prompts: Discovering Rich Visual Clues through Mining Question-Answer Prompts for VQA requiring Diverse World Knowledge*. 2024. arXiv: 2401.10712 `[cs.CV]`.

[124] Taehee Kim, Yeongjae Cho, Heejun Shin, Yohan Jo, and Dongmyung Shin. *Generalizing Visual Question Answering from Synthetic to Human-Written Questions via a Chain of QA with a Large Language Model*. 2024. arXiv: 2401.06400 `[cs.CL]`.

[125] Sadid A. Hasan, Yuan Ling, Oladimeji Farri, Joey Liu, Matthew Lungren, and Henning Müller. "Overview of the ImageCLEF 2018 Medical Domain Visual Question Answering Task". In: *CLEF2018 Working Notes.* CEUR Workshop Proceedings. Avignon, France: CEUR-WS.org <http://ceur-ws.org>, Sept. 2018.

[126] Binh D Nguyen, Thanh-Toan Do, Binh X Nguyen, Tuong Do, Erman Tjiputra, and Quang D Tran. "Overcoming data limitation in medical visual question answering". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention.* Springer. 2019, pp. 522–530.

[127] Li-Ming Zhan, Bo Liu, Lu Fan, Jiaxin Chen, and Xiao-Ming Wu. "Medical visual question answering via conditional reasoning". In: *Proceedings of the 28th ACM International Conference on Multimedia.* 2020, pp. 2345–2354.

[128] Aisha Al-Sadi, Al-Ayyoub M Hana'Al-Theiabat, and Mahmoud Al-Ayyoub. "The Inception Team at VQA-Med 2020: Pretrained VGG with Data Augmentation for Medical VQA and VQG." In: *CLEF (Working Notes).* 2020.

[129] Minh H Vu, Tommy Löfstedt, Tufve Nyholm, and Raphael Sznitman. "A question-centric model for visual question answering in medical imaging". In: *IEEE transactions on medical imaging* 39.9 (2020), pp. 2856–2868.

[130] Haifan Gong, Guanqi Chen, Sishuo Liu, Yizhou Yu, and Guanbin Li. "Cross-Modal Self-Attention with Multi-Task Pre-Training for Medical Visual Question Answering". In: *Proceedings of the 2021 International Conference on Multimedia Retrieval.* 2021, pp. 456–460.

[131] Usman Naseem, Matloob Khushi, and Jinman Kim. "Vision-language transformer for interpretable pathology visual question answering". In: *IEEE Journal of Biomedical and Health Informatics* 27.4 (2022), pp. 1681–1690.

[132] Tom van Sonsbeek, Mohammad Mahdi Derakhshani, Ivona Najdenkoska, Cees G. M. Snoek, and Marcel Worring. *Open-Ended Medical Visual Question Answering Through Prefix Tuning of Language Models.* 2023. arXiv: 2303.05977 [cs.CV].

[133] Lalithkumar Seenivasan, Mobarakol Islam, Gokul Kannan, and Hongliang Ren. "SurgicalGPT: End-to-End Language-Vision GPT for Visual Question Answering in Surgery". In: *arXiv preprint arXiv:2304.09974* (2023).

[134] Jinlong He, Pengfei Li, Gang Liu, Zixu Zhao, and Shenjun Zhong. *PeFoMed: Parameter Efficient Fine-tuning on Multimodal Large Language Models for Medical Visual Question Answering.* 2024. arXiv: 2401.02797 [cs.CL].

[135] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. "Making the v in vqa matter: Elevating the role of image understanding in visual question answering". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 2017, pp. 6904–6913.

[136] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. "Don't just assume; look and answer: Overcoming priors for visual question answering". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 4971–4980.

[137] Asma Ben Abacha, Mourad Sarrouti, Dina Demner-Fushman, Sadid A. Hasan, and Henning Müller. "Overview of the VQA-Med Task at ImageCLEF 2021: Visual Question Answering and Generation in the Medical Domain". In: *CLEF 2021 Working Notes*. CEUR Workshop Proceedings. Bucharest, Romania: CEUR-WS.org, Sept. 2021.

[138] Jie Zhu, Ellean Zhang, and Katia Del Rio-Tsonis. "Eye anatomy". In: *eLS* (2012).

[139] Zhenguo Yang, Jiale Xiang, Jiuxiang You, Qing Li, and Wenyin Liu. "Event-Oriented Visual Question Answering: The E-VQA Dataset and Benchmark". In: *IEEE Transactions on Knowledge and Data Engineering* (2023).

[140] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. *OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge*. 2019. arXiv: 1906.00067 `[cs.CV]`.

[141] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. "Towards vqa models that can read". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 8317–8326.

[142] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. "Docvqa: A dataset for vqa on document images". In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2021, pp. 2200–2209.

[143] Jingying Gao, Qi Wu, Alan Blair, and Maurice Pagnucco. "Lora: A logical reasoning augmented dataset for visual question answering". In: *Advances in Neural Information Processing Systems* 36 (2024).

[144] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. "Visual7w: Grounded question answering in images". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 4995–5004.

[145] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2901–2910.

[146] Olga Kovaleva et al. "Towards visual dialog for radiology". In: *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*. 2020, pp. 60–69.

[147] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. "A dataset of clinically generated visual questions and answers about radiology images". In: *Scientific data* 5.1 (2018), pp. 1–10.

[148] Asma Ben Abacha, Vivek V. Datla, Sadid A. Hasan, Dina Demner-Fushman, and Henning Müller. "Overview of the VQA-Med Task at ImageCLEF 2020: Visual Question Answering and Generation in the Medical Domain". In: *CLEF 2020 Working Notes*. CEUR Workshop Proceedings. Thessaloniki, Greece: CEUR-WS.org, Sept. 2020.

[149] Sergio Tascon-Morales, Pablo Márquez-Neila, and Raphael Sznitman. "Consistency-Preserving Visual Question Answering in Medical Imaging". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VIII*. Springer. 2022, pp. 386–395.

[150] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. "Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering". In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2021, pp. 1650–1654.

[151] Asma Ben Abacha, Sadid A. Hasan, Vivek V. Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. "VQA-Med: Overview of the Medical Visual Question Answering Task at ImageCLEF 2019". In: *CLEF2019 Working Notes*. CEUR Workshop Proceedings. Lugano, Switzerland: CEUR-WS.org <http://ceur-ws.org>, Sept. 2019.

[152] Bogdan Ionescu, Henning Müller, Ana-Maria Druagulinescu, Wen wai Yim, Asma Ben Abacha, Neal Snider, et al. "Overview of ImageCLEF 2023: Multimedia Retrieval in Medical, SocialMedia and Recommender Systems Applications". In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Proceedings of the 14th International Conference of the CLEF Association (CLEF 2023). Thessaloniki, Greece: Springer Lecture Notes in Computer Science LNCS, Sept. 2023.

[153] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. "Pathvqa: 30000+ questions for medical visual question answering". In: *arXiv preprint arXiv:2003.10286* (2020).

[154] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. "Pmc-vqa: Visual instruction tuning for medical visual question answering". In: *arXiv preprint arXiv:2305.10415* (2023).

[155] Sight Research UK. *How Your Eyes Work — sightresearchuk.org*. https://www.sightresearchuk.org/how-your-eyes-work/. [Accessed 16-02-2024].

[156] Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, and Fabrice Meriaudeau. "Indian Diabetic Retinopathy Image Dataset (IDRiD)". In: (2018). DOI: 10.21227/H25W98.

[157] Rui Bernardes, Pedro Serranho, and Conceição Lobo. "Digital ocular fundus imaging: a review". In: *Ophthalmologica* 226.4 (2011), pp. 161–181.

[158] Fulong Ren, Peng Cao, Dazhe Zhao, and Chao Wan. "Diabetic macular edema grading in retinal images using vector quantization and semi-supervised learning". In: *Technology and Health Care* 26.S1 (2018), pp. 389–397.

[159] Sergio Tascon-Morales, Pablo Márquez-Neila, and Raphael Sznitman. “Localized Questions in Medical Visual Question Answering”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 361–370.

[160] Yonglin Yu, Haifeng Li, Hanrong Shi, Lin Li, and Jun Xiao. “Question-guided feature pyramid network for medical visual question answering”. In: *Expert Systems with Applications* 214 (2023), p. 119148.

[161] Fuji Ren and Yangyang Zhou. “Cgmvqa: A new classification and generative model for medical visual question answering”. In: *IEEE Access* 8 (2020), pp. 50626–50636.

[162] Binh D. Nguyen, Thanh-Toanidrid Do, Binh X. Nguyen, Tuong Do, Erman Tjiputra, and Quang D. Tran. “Overcoming Data Limitation in Medical Visual Question Answering”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Ed. by Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan. Cham: Springer International Publishing, 2019, pp. 522–530. ISBN: 978-3-030-32251-9.

[163] Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and Christoph M Friedrich. “Radiology Objects in COntext (ROCO): a multimodal image dataset”. In: *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3*. Springer. 2018, pp. 180–189.

[164] Tuong Do, Binh X Nguyen, Erman Tjiputra, Minh Tran, Quang D Tran, and Anh Nguyen. “Multiple meta-model quantifying for medical visual question answering”. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*. Springer. 2021, pp. 64–74.

[165] Arjun Mani, Nobline Yoo, Will Hinthorn, and Olga Russakovsky. “Point and ask: Incorporating pointing into visual question answering”. In: *arXiv preprint arXiv:2011.13681* (2020).

[166] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. “Hadamard product for low-rank bilinear pooling”. In: *arXiv preprint arXiv:1610.04325* (2016).

[167] Max Allan et al. “2017 robotic instrument segmentation challenge”. In: *arXiv preprint arXiv:1902.06426* (2019).

[168] Markus Fox, Mario Taschwer, and Klaus Schoeffmann. “Pixel-Based Tool Segmentation in Cataract Surgery Videos with Mask R-CNN”. In: *33rd IEEE International Symposium on Computer-Based Medical Systems, CBMS 2020, Rochester, MN, USA, July 28-30, 2020*. Ed. by Alba García Seco de Herrera, Alejandro Rodríguez González, K. C. Santosh, Zelalem Temesgen, Bridget Kane, and Paolo Soda. IEEE, 2020, pp. 565–568. DOI: 10.1109/CBMS49503.2020.00112.

[169] Duzhen Zhang, Yahan Yu, Chenxing Li, Jiahua Dong, Dan Su, Chenhui Chu, and Dong Yu. "Mm-llms: Recent advances in multimodal large language models". In: *arXiv preprint arXiv:2401.13601* (2024).

[170] Marco Tulio Ribeiro, Carlos Guestrin, and Sameer Singh. "Are red roses red? evaluating consistency of question-answering models". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, pp. 6174–6184.

[171] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. "Multimodal few-shot learning with frozen language models". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 200–212.

[172] Hugo Touvron et al. "Llama 2: Open foundation and fine-tuned chat models". In: *arXiv preprint arXiv:2307.09288* (2023).

[173] Mourad Sarrouti. "NLM at VQA-Med 2020: Visual Question Answering and Generation in the Medical Domain." In: *CLEF (Working Notes)*. 2020.

[174] Peiqi Wang, Ruizhi Liao, Daniel Moyer, Seth Berkowitz, Steven Horng, and Polina Golland. "Image Classification with Consistent Supporting Evidence". In: *Machine Learning for Health*. PMLR. 2021, pp. 168–180.

[175] Vatsal Goel, Mohit Chandak, Ashish Anand, and Prithwijit Guha. "IQ-VQA: Intelligent Visual Question Answering". In: *International Conference on Pattern Recognition*. Springer. 2021, pp. 357–370.

[176] Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. "VQA-LOL: Visual question answering under the lens of logic". In: *European conference on computer vision*. Springer. 2020, pp. 379–396.

[177] Arijit Ray, Karan Sikka, Ajay Divakaran, Stefan Lee, and Giedrius Burachas. "Sunny and dark outside?! improving answer consistency in vqa through entailed question generation". In: *arXiv preprint arXiv:1909.04696* (2019).

[178] Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. "Cycle-consistency for robust visual question answering". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 6649–6658.

[179] Damien Teney, Ehsan Abbasnejad, and Anton van den Hengel. "On incorporating semantic prior knowledge in deep learning through embedding-space constraints". In: *arXiv preprint arXiv:1909.13471* (2019).

[180] Yuanyuan Yuan, Shuai Wang, Mingyue Jiang, and Tsong Yueh Chen. "Perception Matters: Detecting Perception Failures of VQA Models Using Metamorphic Testing". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 16908–16917.

[181] Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. "Rubi: Reducing unimodal biases for visual question answering". In: *Advances in neural information processing systems* 32 (2019), pp. 841–852.

[182] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. “Show, attend and tell: Neural image caption generation with visual attention”. In: *International conference on machine learning*. PMLR. 2015, pp. 2048–2057.

[183] Etienne Decenciere et al. “TeleOphta: Machine learning and image processing methods for teleophthalmology”. In: *Irbm* 34.2 (2013), pp. 196–203.

[184] Sergio Tascon-Morales, Pablo Márquez-Neila, and Raphael Sznitman. “Logical Implications for Visual Question Answering Consistency”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 6725–6735.

[185] Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel, and Anthony Dick. “Explicit knowledge-based reasoning for visual question answering”. In: *arXiv preprint arXiv:1511.02570* (2015).

[186] Remi Cadene, Hedi Ben-Younes, Matthieu Cord, and Nicolas Thome. “Murel: Multimodal relational reasoning for visual question answering”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 1989–1998.

[187] Yirui Wu, Yuntao Ma, and Shaohua Wan. “Multi-scale relation reasoning for multimodal Visual Question Answering”. In: *Signal Processing: Image Communication* 96 (2021), p. 116319.

[188] Chenchen Jing, Yunde Jia, Yuwei Wu, Xinyu Liu, and Qi Wu. “Maintaining Reasoning Consistency in Compositional Visual Question Answering”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 5099–5108.

[189] Vedika Agarwal, Rakshith Shetty, and Mario Fritz. “Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 9690–9698.

[190] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. “Counterfactual samples synthesizing for robust visual question answering”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 10800–10809.

[191] Ehsan Abbasnejad, Damien Teney, Amin Parvaneh, Javen Shi, and Anton van den Hengel. “Counterfactual vision and language learning”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 10044–10054.

[192] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. “Dual attention networks for multimodal reasoning and matching”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 299–307.

[193] Corentin Dancette, Remi Cadene, Damien Teney, and Matthieu Cord. “Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 1574–1583.

[194] Xinzhe Han, Shuhui Wang, Chi Su, Qingming Huang, and Qi Tian. “Greedy gradient ensemble for robust visual question answering”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 1584–1593.

[195] Pratyay Banerjee, Tejas Gokhale, Yezhou Yang, and Chitta Baral. “Weakly Supervised Relative Spatial Reasoning for Visual Question Answering”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 1908–1918.

[196] Qingxing Cao, Wentao Wan, Keze Wang, Xiaodan Liang, and Liang Lin. “Linguistically routing capsule network for out-of-distribution visual question answering”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 1614–1623.

[197] Xu Yang, Chongyang Gao, Hanwang Zhang, and Jianfei Cai. “Auto-parsing network for image captioning and visual question answering”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 2197–2207.

[198] Yiyi Zhou, Tianhe Ren, Chaoyang Zhu, Xiaoshuai Sun, Jianzhuang Liu, Xinghao Ding, Mingliang Xu, and Rongrong Ji. “Trar: Routing the attention spans in transformer for visual question answering”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 2074–2084.

[199] Yang Ding, Jing Yu, Bang Liu, Yue Hu, Mingxin Cui, and Qi Wu. “MuKEA: Multimodal Knowledge Extraction and Accumulation for Knowledge-based Visual Question Answering”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 5089–5098.

[200] Feng Gao, Qing Ping, Govind Thattai, Aishwarya Reganti, Ying Nian Wu, and Prem Natarajan. “Transform-Retrieve-Generate: Natural Language-Centric Outside-Knowledge Visual Question Answering”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 5067–5077.

[201] Vipul Gupta, Zhuowan Li, Adam Kortylewski, Chenyu Zhang, Yingwei Li, and Alan Yuille. “Swapmix: Diagnosing and regularizing the over-reliance on visual context in visual question answering”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 5078–5088.

[202] Matthew Walmer, Karan Sikka, Indranil Sur, Abhinav Shrivastava, and Susmit Jha. “Dual-Key Multimodal Backdoors for Visual Question Answering”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 15375–15385.

[203] Chongyan Chen, Samreen Anjum, and Danna Gurari. “Grounding Answers for Visual Questions Asked by Visually Impaired People”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 19098–19107.

[204] Bill MacCartney and Christopher D Manning. “Modeling semantic containment and exclusion in natural language inference”. In: *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*. 2008, pp. 521–528.

[205] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. “From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions”. In: *Transactions of the Association for Computational Linguistics* 2 (2014), pp. 67–78.

[206] Adina Williams, Nikita Nangia, and Samuel R Bowman. “A broad-coverage challenge corpus for sentence understanding through inference”. In: *arXiv preprint arXiv:1704.05426* (2017).

[207] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. “Superglue: A stickier benchmark for general-purpose language understanding systems”. In: *Advances in neural information processing systems* 32 (2019).

[208] Aalok Sathe, Salar Ather, Tuan Manh Le, Nathan Perry, and Joonsuk Park. “Automated fact-checking of claims from wikipedia”. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. 2020, pp. 6874–6882.

[209] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. “Adversarial NLI: A new benchmark for natural language understanding”. In: *arXiv preprint arXiv:1910.14599* (2019).

[210] Yinhan Liu et al. “Roberta: A robustly optimized bert pretraining approach”. In: *arXiv preprint arXiv:1907.11692* (2019).

[211] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. “Deberta: Decoding-enhanced bert with disentangled attention”. In: *arXiv preprint arXiv:2006.03654* (2020).

[212] R. Bradley and N. Swartz. *Possible Worlds: An Introduction to Logic and Its Philosophy*. B. Blackwell, 1979. ISBN: 9780631161400.

[213] Slav Petrov, Dipanjan Das, and Ryan McDonald. “A universal part-of-speech tagset”. In: *arXiv preprint arXiv:1104.2086* (2011).

[214] Weijie Su. *Pythia*. https://github.com/jackroos/pythia. 2019.

[215] Yiqi Wang et al. *Exploring the Reasoning Abilities of Multimodal Large Language Models (MLLMs): A Comprehensive Survey on Emerging Trends in Multimodal Reasoning*. 2024. arXiv: 2401.06805 [cs.CL].

[216] Jingxuan Wei, Cheng Tan, Zhangyang Gao, Linzhuang Sun, Siyuan Li, Bihui Yu, Ruifeng Guo, and Stan Z. Li. *Enhancing Human-like Multi-Modal Reasoning: A New Challenging Dataset and Comprehensive Framework*. 2023. arXiv: 2307.12626 [cs.AI].

[217] Yue Fan, Jing Gu, Kaiwen Zhou, Qianqi Yan, Shan Jiang, Ching-Chen Kuo, Xinze Guan, and Xin Eric Wang. “Muffin or Chihuahua? Challenging Large Vision-Language Models with Multipanel VQA”. In: *arXiv preprint arXiv:2401.15847* (2024).

[218] Bo Peng et al. *RWKV: Reinventing RNNs for the Transformer Era*. 2023. arXiv: 2305.13048 [cs.CL].

[219] Wenhui Wang et al. *Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks*. 2022. arXiv: 2208.10442 [cs.CV].

[220] Chantal Pellegrini, Ege Özsoy, Benjamin Busam, Nassir Navab, and Matthias Keicher. "RaDialog: A Large Vision-Language Model for Radiology Report Generation and Conversational Assistance". In: *arXiv preprint arXiv:2311.18681* (2023).

[221] Ming Y Lu et al. "A Foundational Multimodal Vision Language AI Assistant for Human Pathology". In: *arXiv preprint arXiv:2312.07814* (2023).

[222] Yuxuan Sun et al. "Pathasst: Redefining pathology through generative foundation ai assistant for pathology". In: *arXiv preprint arXiv:2305.15072* (2023).

A Localized Questions in Medical Visual Question Answering

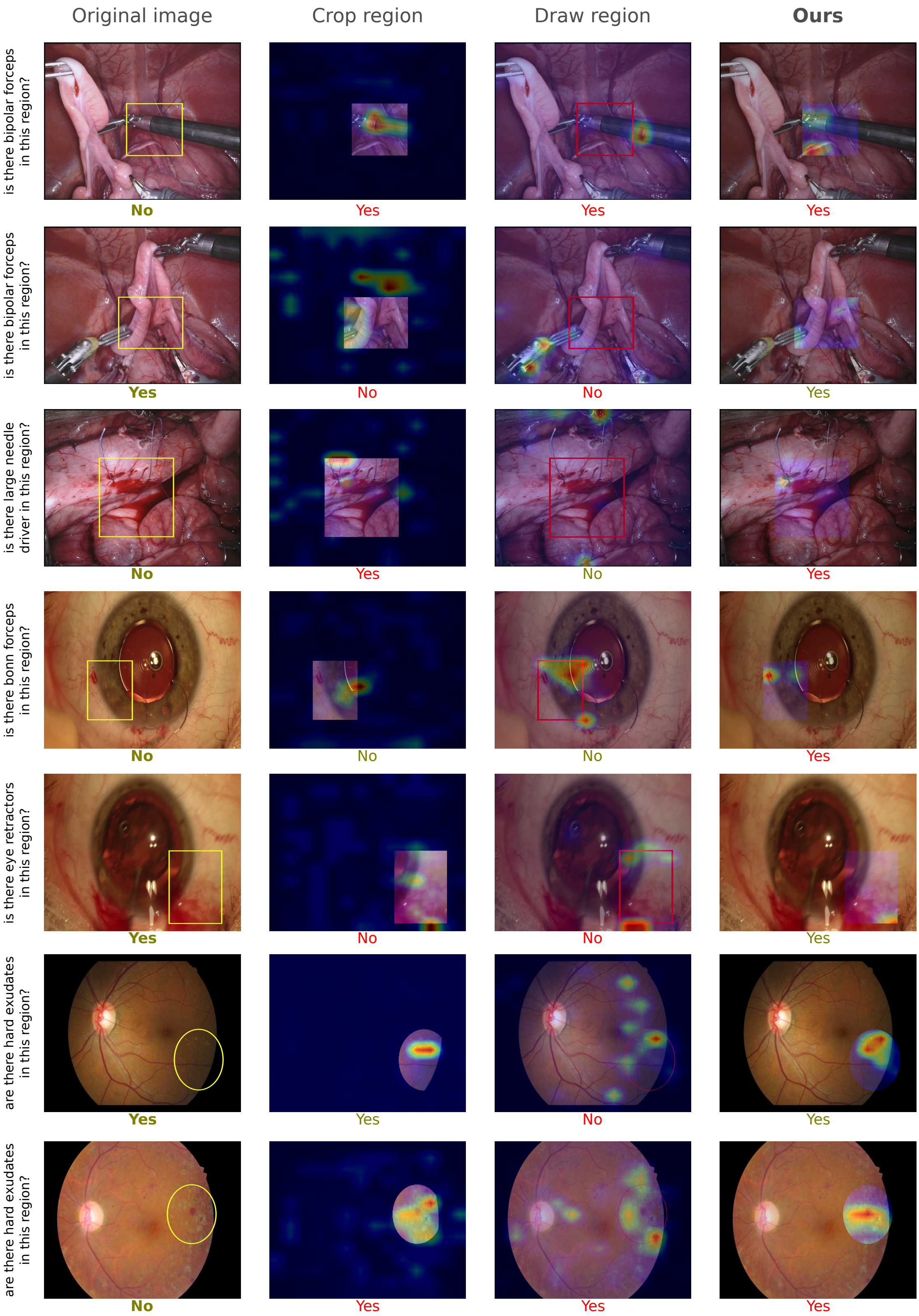

Original image
Crop region
Draw region
Ours
is there bipolar forceps in this region?
No
Yes
Yes
Yes
is there bipolar forceps in this region?
Yes
No
No
Yes
is there large needle driver in this region?
No
Yes
No
Yes
is there bonn forceps in this region?
No
No
No
Yes
is there eye retractors in this region?
Yes
No
No
Yes
are there hard exudates in this region?
Yes
Yes
No
Yes
are there hard exudates in this region?
No
Yes
Yes
Yes

FIGURE A.1: Additional qualitative examples from the RIS-VQA (rows 1-3), INSEGCAT-VQA (rows 4-5) and DME-VQA (last two rows) datasets. The first column shows the image, the region, and the ground truth answer. Other columns show the overlaid attention maps and the answers produced by each model. Wrong answers are shown in red.

Instrument	Method				
	Ignore mask	Region in Text	Crop Region	Draw Region	**Ours**
Eye retractors	0.500 ±0	0.822 ±0.005	0.882 ±0.002	**0.913** ±**0.002**	0.912 ±0.004
Rycroft cannula	0.500 ±0	0.728 ±0.025	0.847 ±0.002	0.910 ±0.002	**0.912** ±**0.001**
Viter. handpiece	0.500 ±0	0.805 ±0.024	**0.960** ±**0.006**	0.953 ±0.005	0.881 ±0.010
Secondary knife	0.500 ±0	0.751 ±0.013	0.822 ±0.013	0.828 ±0.011	**0.830** ±**0.007**
Suture needle	0.500 ±0	0.825 ±0.019	**0.889** ±**0.004**	0.851 ±0.014	0. 848 ±0.018
Micro-manipulator	0.500 ±0	0.645 ±0.032	0.846 ±0.003	0.877 ±0.007	**0.889** ±**0.004**
Bonn forceps	0.500 ±0	0.846 ±0.020	0.908 ±0.008	0.897 ±0.012	**0.909** ±**0.004**
Visco. cannula	0.500 ±0	0.879 ±0.013	**0.954** ±**0.002**	0.946 ±0.003	0.937 ±0.003
Cap. forceps	0.500 ±0	0.804 ±0.039	0.905 ±0.010	**0.932** ±**0.018**	0.908 ±0.010
Phaco. handpiece	0.500 ±0	0.717 ±0.037	0.900 ±0.008	0.920 ±0.003	**0.952** ±**0.003**
Charleux cannula	0.500 ±0	0.841 ±0.044	0.826 ±0.029	0.915 ±0.016	**0.925** ±**0.016**
Lens injector	0.500 ±0	0.773 ±0.006	**0.941** ±**0.003**	0.927 ±0.007	0.917 ±0.007
Cap. cystotome	0.500 ±0	0.789 ±0.009	0.938 ±0.006	0.934 ±0.002	**0.953** ±**0.001**
Primary knife	0.500 ±0	0.846 ±0.011	**0.954** ±**0.003**	0.926 ±0.008	0.941 ±0.004
Hydro. cannula	0.500 ±0	0.865 ±0.006	0.939 ±0.004	**0.945** ±**0.006**	0.940 ±0.004
A/I handpiece	0.500 ±0	0.877 ±0.021	0.935 ±0.006	0.906 ±0.004	**0.938** ±**0.001**

TABLE A.1: Average test AUC for different methods on INSEGCAT-VQA.

B Targeted Visual Prompting for Medical Visual Question Answering

Method	Accuracy (%)			
	Overall	Grade	Whole	Macula
No Mask	60.50	81.13	76.42	85.85
Region in Text	64.75	79.25	83.96	82.08
Crop Region	86.05	80.19	83.96	84.91
Draw Region	86.18	79.25	83.02	83.02
Context Only	82.61	76.42	87.74	90.57
Ours	89.29	79.25	83.96	84.91

TABLE B.1: Accuracy for the DME-VQA dataset by question type.

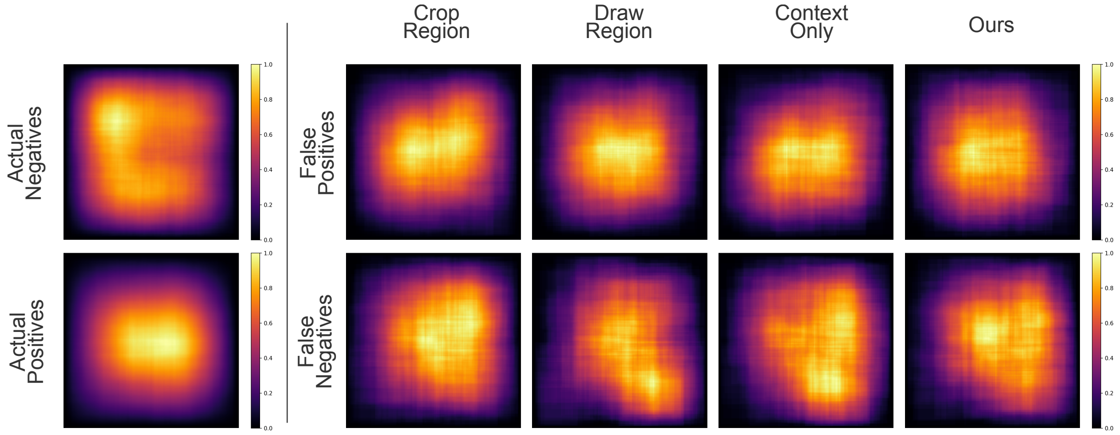


FIGURE B.1: Error analysis by region location for the four strongest baselines for the RIS-VQA dataset. The maps are obtained by adding binary masks representing the regions for all QA pairs in each category and then normalizing.

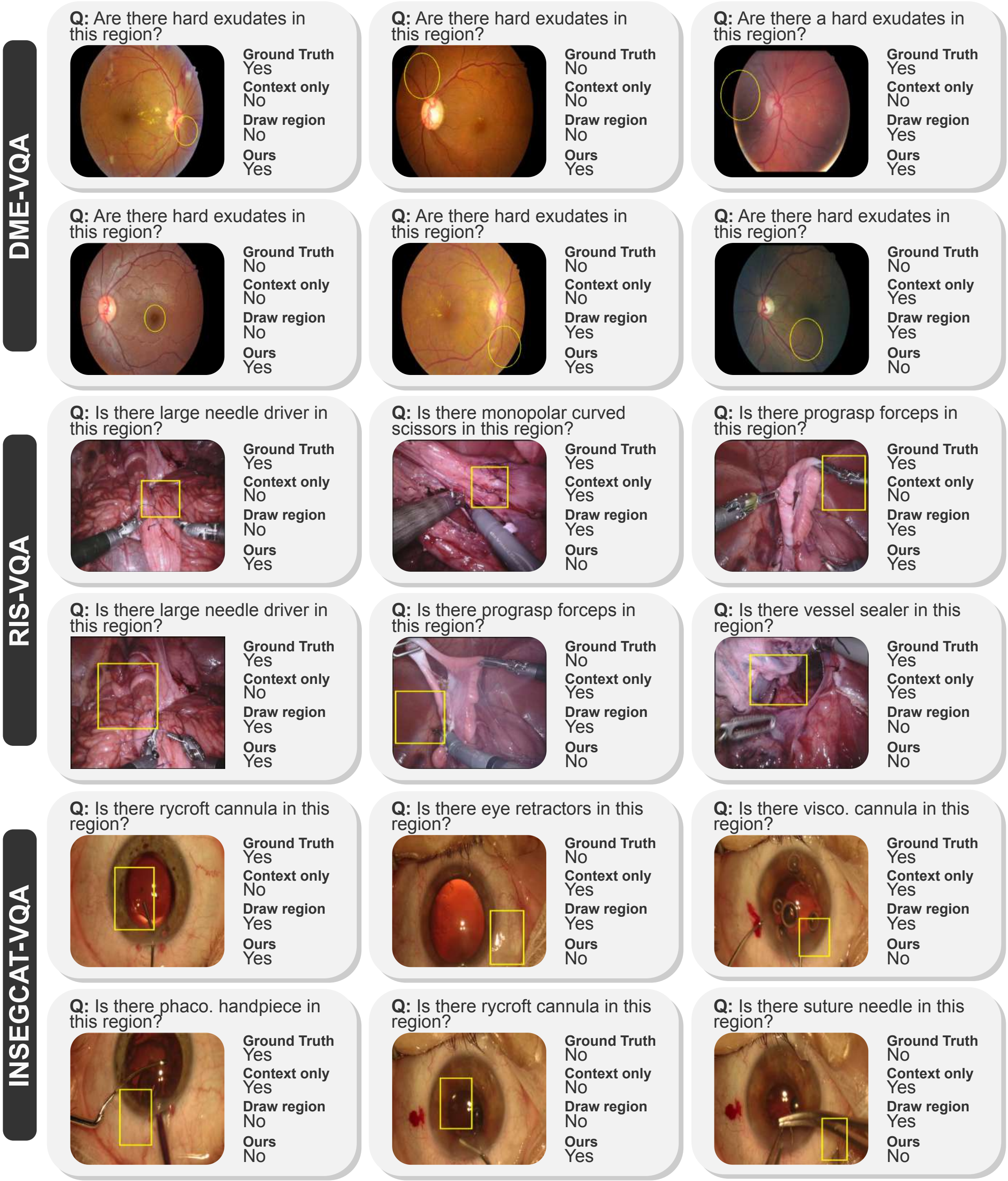


FIGURE B.2: Additional examples for DME-VQA (rows 1 and 2), RIS-VQA (rows 3 and 4) and Insegcat-VQA (rows 5 and 6).

C Consistency-preserving Visual Question Answering in Medical Imaging

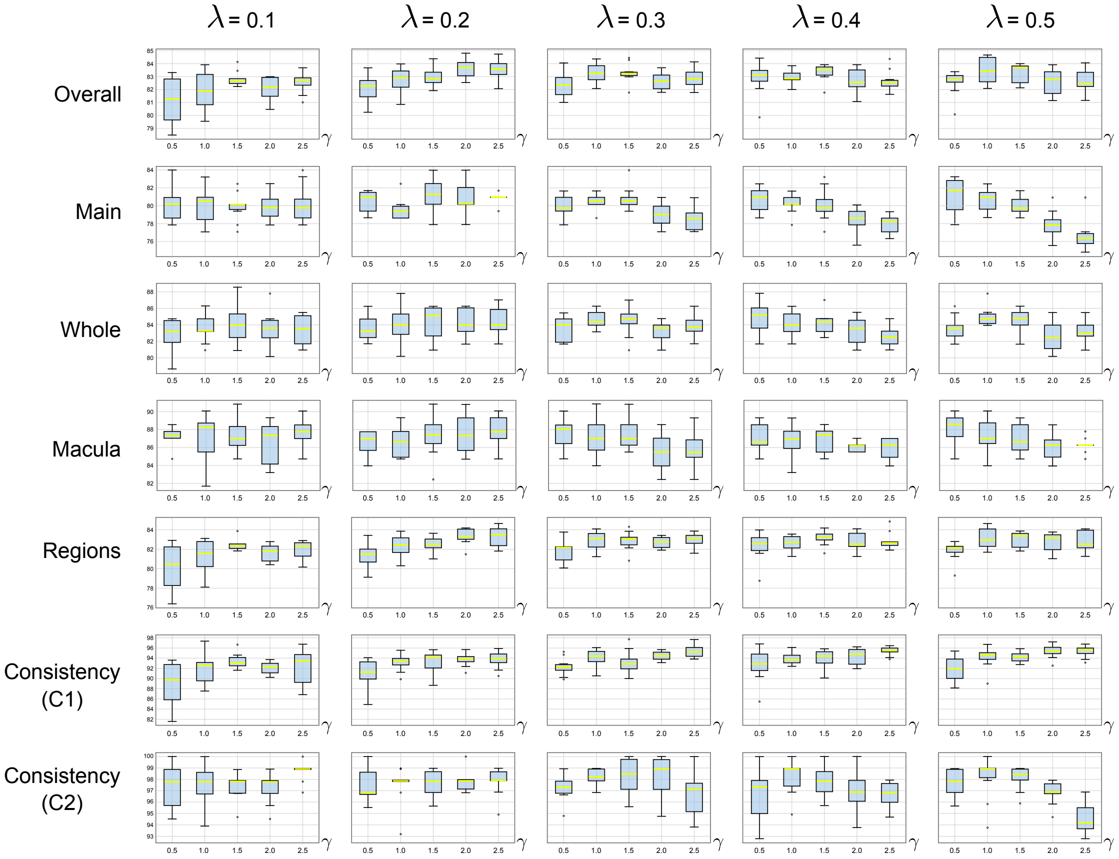
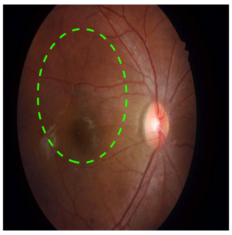
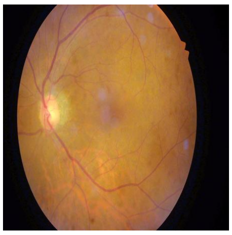


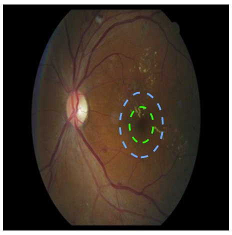
FIGURE C.1: Effect of the variation of the hyperparameters λ and γ, for each metric. The first 5 rows refer to accuracy for all questions (overall), for main questions (main) and for sub-questions (whole, macula and regions). The last two rows correspond to the consistency. In general, a higher value of λ leads to a higher consistency, which is the expected behavior. High values of both parameters can produce a decrease in the accuracy of main questions.



Question	Type	Ans. GT	Ans. baseline	Ans. SQuINT	Ans. Ours
What is the DME grade?	main	0	0	0	0
Are there hard exudates in the image?	sub	NO	YES	NO	NO
Are there hard exudates in the macula?	sub	NO	NO	NO	NO
Are there hard exudates in this region?	sub	NO	YES	YES	NO

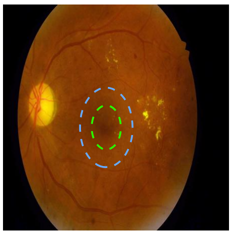


Question	Type	Ans. GT	Ans. baseline	Ans. SQuINT	Ans. Ours
What is the DME grade?	main	0	0	0	0
Are there hard exudates in the image?	sub	NO	NO	YES	NO
Are there hard exudates in the macula?	sub	NO	NO	NO	NO



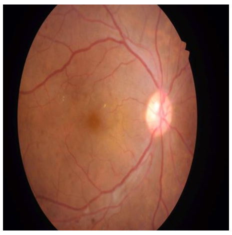
Question	Type	Ans. GT	Ans. baseline	Ans. SQuINT	Ans. Ours
What is the DME grade?	main	2	2	2	2
Are there hard exudates in the image?	sub	YES	YES	YES	YES
Are there hard exudates in the macula?	sub	YES	YES	YES	YES
Are there hard exudates in this region?	sub	YES	NO	NO	NO
Are there hard exudates in this region?	sub	YES	YES	YES	YES

*Regions located at fovea center, with radius smaller than 1 optic disc diameter (See Fig. 3)

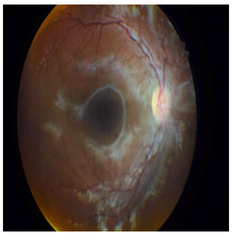


Question	Type	Ans. GT	Ans. baseline	Ans. SQuINT	Ans. Ours
What is the DME grade?	main	2	2	2	2
Are there hard exudates in the image?	sub	YES	YES	YES	YES
Are there hard exudates in the macula?	sub	YES	YES	YES	YES
Are there hard exudates in this region?	sub	YES	NO	YES	NO
Are there hard exudates in this region?	sub	YES	YES	YES	NO

*Regions located at fovea center, with radius smaller than 1 optic disc diameter (See Fig. 3)



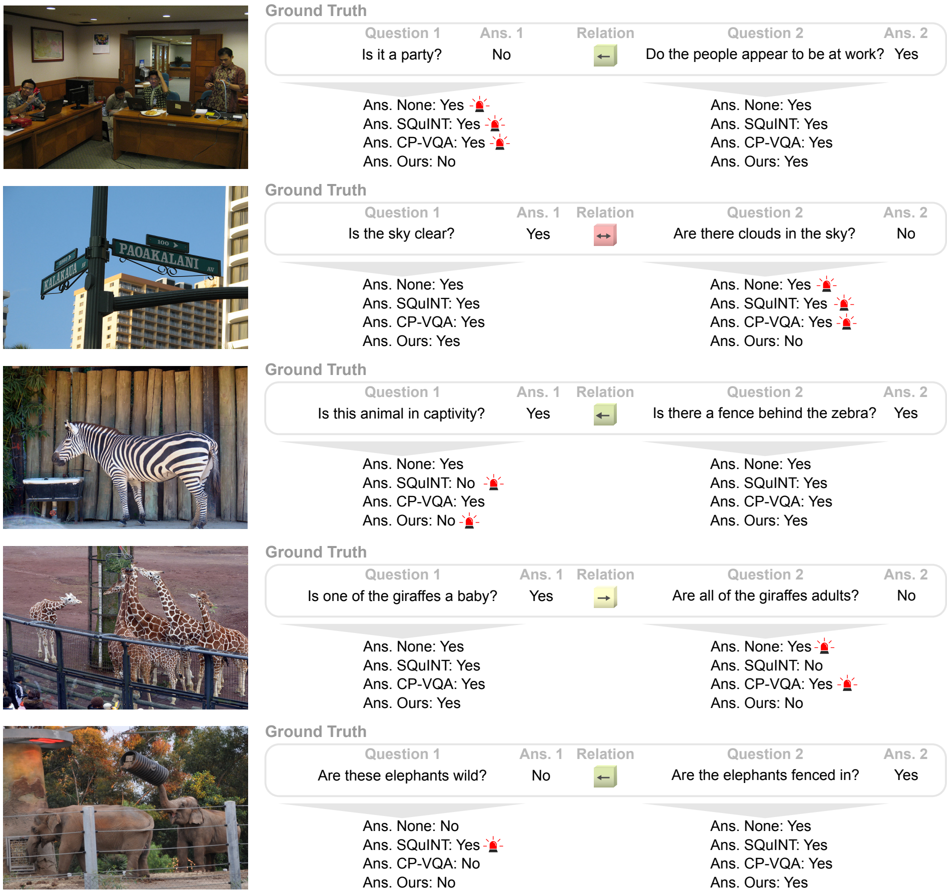
Question	Type	Ans. GT	Ans. baseline	Ans. SQuINT	Ans. Ours
What is the DME grade?	main	2	2	0	2
Are there hard exudates in the image?	sub	YES	YES	NO	YES
Are there hard exudates in the macula?	sub	YES	YES	NO	YES



Question	Type	Ans. GT	Ans. baseline	Ans. SQuINT	Ans. Ours
What is the DME grade?	main	0	2	2	0
Are there hard exudates in the image?	sub	NO	YES	YES	NO
Are there hard exudates in the macula?	sub	NO	YES	YES	NO

FIGURE C.2: Additional qualitative examples from the DME dataset. Inconsistent answers are highlighted in red. A more consistent behavior is observed in our method in comparison to the baselines (rows 1-2). Even though our method can make mistakes (rows 3-4), it also shows an improvement in the performance on main questions (rows 5-6).

D Logical Implications for Visual Question Answering Consistency



Ground Truth

Question 1	Ans. 1	Relation	Question 2	Ans. 2
Is it a party?	No	←	Do the people appear to be at work?	Yes

Ans. None: Yes 🚨	Ans. None: Yes
Ans. SQuINT: Yes 🚨	Ans. SQuINT: Yes
Ans. CP-VQA: Yes 🚨	Ans. CP-VQA: Yes
Ans. Ours: No	Ans. Ours: Yes

Ground Truth

Question 1	Ans. 1	Relation	Question 2	Ans. 2
Is the sky clear?	Yes	↔	Are there clouds in the sky?	No

Ans. None: Yes	Ans. None: Yes 🚨
Ans. SQuINT: Yes	Ans. SQuINT: Yes 🚨
Ans. CP-VQA: Yes	Ans. CP-VQA: Yes 🚨
Ans. Ours: Yes	Ans. Ours: No

Ground Truth

Question 1	Ans. 1	Relation	Question 2	Ans. 2
Is this animal in captivity?	Yes	←	Is there a fence behind the zebra?	Yes

Ans. None: Yes	Ans. None: Yes
Ans. SQuINT: No 🚨	Ans. SQuINT: Yes
Ans. CP-VQA: Yes	Ans. CP-VQA: Yes
Ans. Ours: No 🚨	Ans. Ours: Yes

Ground Truth

Question 1	Ans. 1	Relation	Question 2	Ans. 2
Is one of the giraffes a baby?	Yes	→	Are all of the giraffes adults?	No

Ans. None: Yes	Ans. None: Yes 🚨
Ans. SQuINT: Yes	Ans. SQuINT: No
Ans. CP-VQA: Yes	Ans. CP-VQA: Yes 🚨
Ans. Ours: Yes	Ans. Ours: No

Ground Truth

Question 1	Ans. 1	Relation	Question 2	Ans. 2
Are these elephants wild?	No	←	Are the elephants fenced in?	Yes

Ans. None: No	Ans. None: Yes
Ans. SQuINT: Yes 🚨	Ans. SQuINT: Yes
Ans. CP-VQA: No	Ans. CP-VQA: Yes
Ans. Ours: No	Ans. Ours: Yes

FIGURE D.1: Additional qualitative examples from the Introspect dataset using BAN as the backbone. Red siren symbols indicate inconsistent cases.

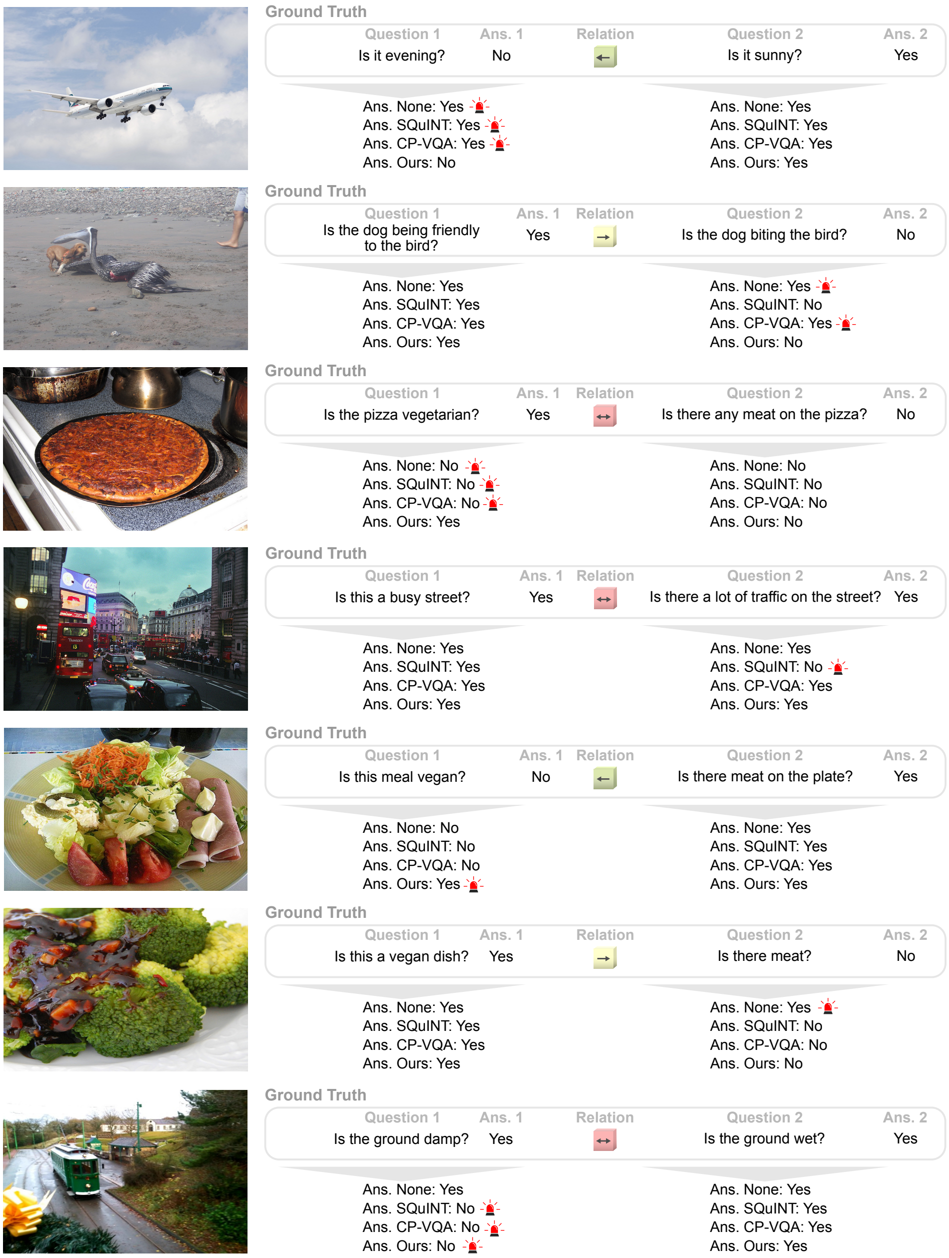


FIGURE D.2: Additional qualitative examples from the Introspect dataset using BAN as the backbone. Red siren symbols indicate inconsistent cases.

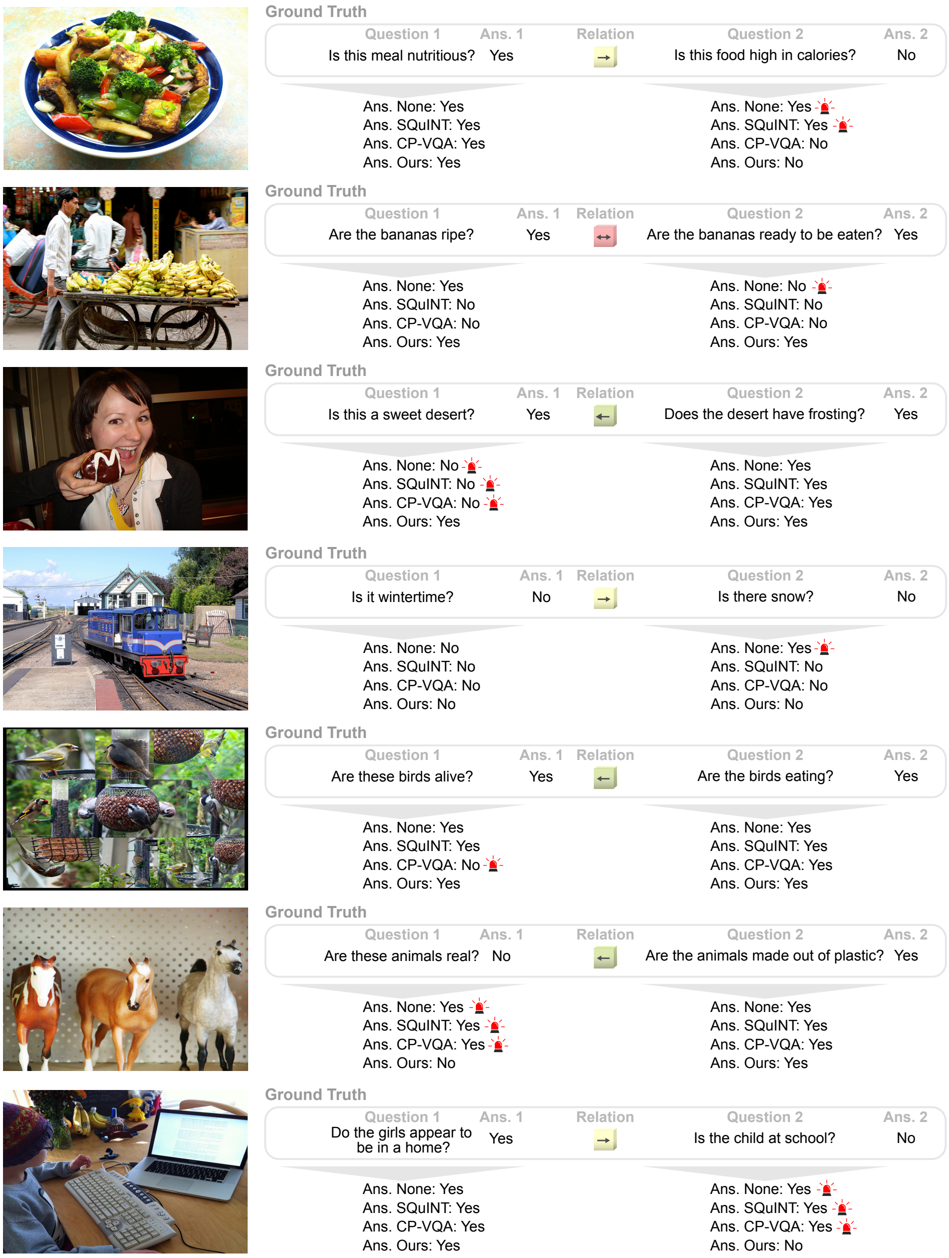


FIGURE D.3: Additional qualitative examples from the Introspect dataset using BAN as the backbone. Red siren symbols indicate inconsistent cases.

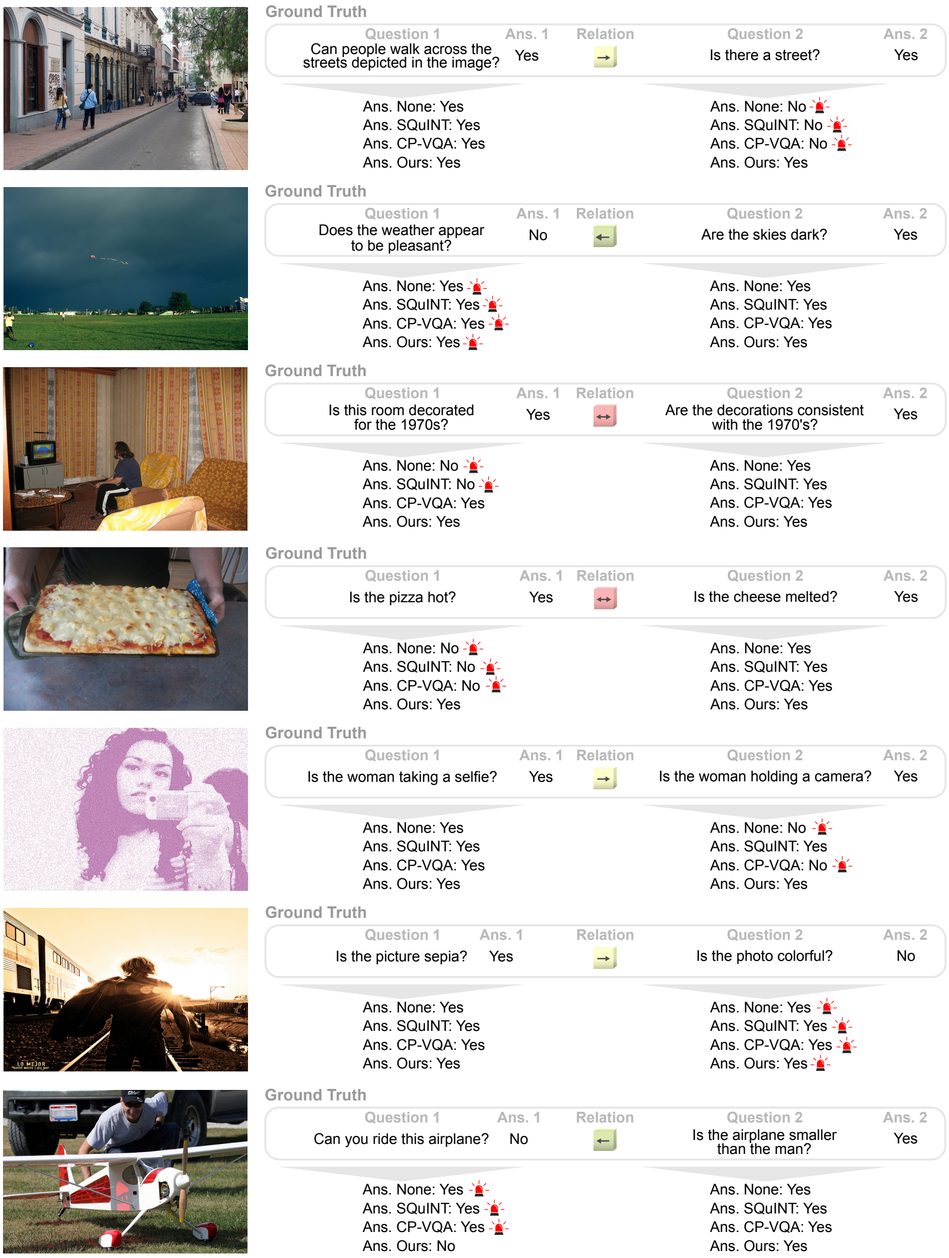
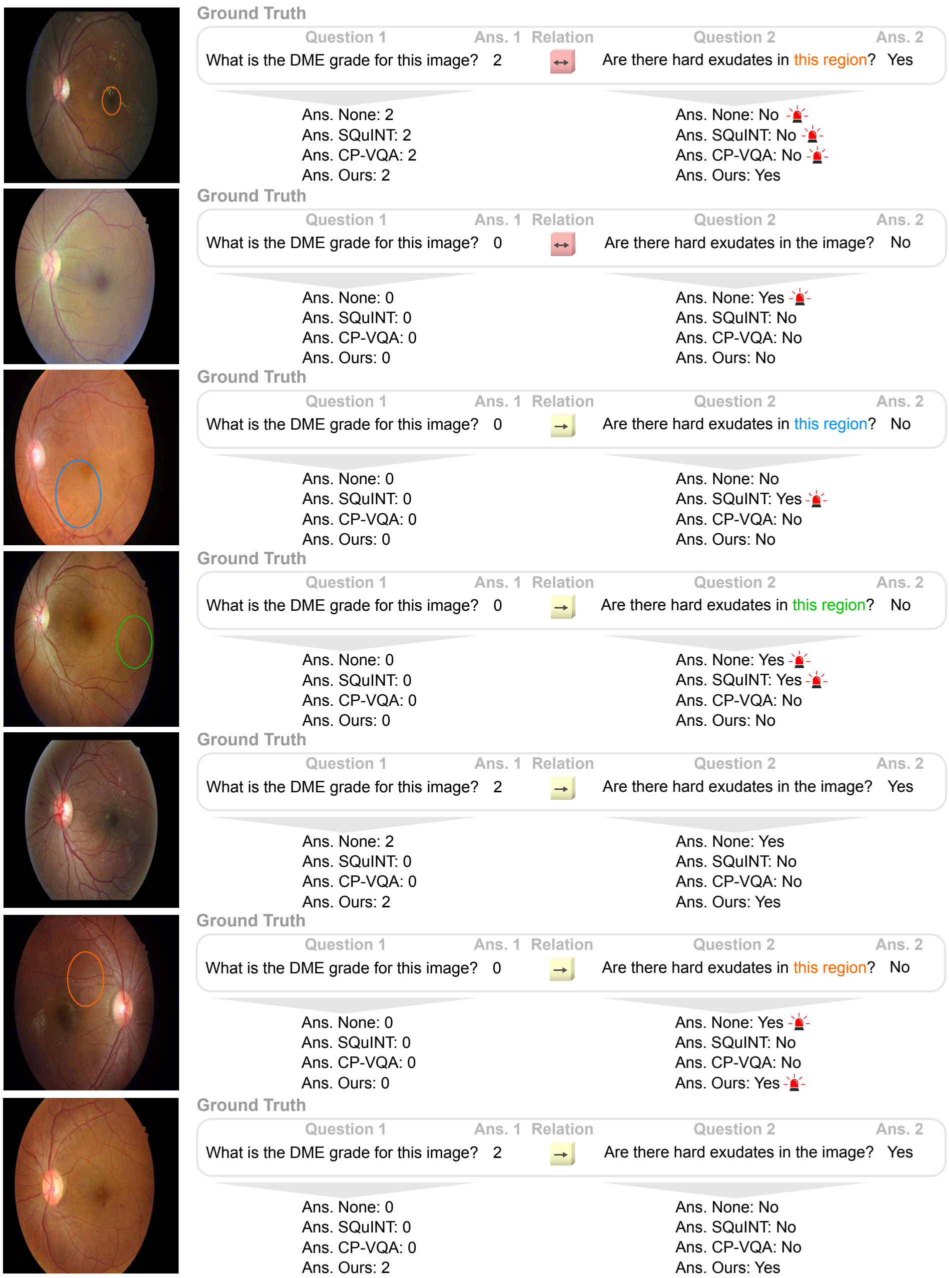


FIGURE D.4: Additional qualitative examples from the Introspect dataset using BAN as the backbone. Red siren symbols indicate inconsistent cases.



Ground Truth

Question 1	Ans. 1	Relation	Question 2	Ans. 2
What is the DME grade for this image?	2	↔	Are there hard exudates in this region?	Yes

Ans. None: 2
Ans. SQuINT: 2
Ans. CP-VQA: 2
Ans. Ours: 2

Ans. None: No 🚨
Ans. SQuINT: No 🚨
Ans. CP-VQA: No 🚨
Ans. Ours: Yes

Ground Truth

Question 1	Ans. 1	Relation	Question 2	Ans. 2
What is the DME grade for this image?	0	↔	Are there hard exudates in the image?	No

Ans. None: 0
Ans. SQuINT: 0
Ans. CP-VQA: 0
Ans. Ours: 0

Ans. None: Yes 🚨
Ans. SQuINT: No
Ans. CP-VQA: No
Ans. Ours: No

Ground Truth

Question 1	Ans. 1	Relation	Question 2	Ans. 2
What is the DME grade for this image?	0	→	Are there hard exudates in this region?	No

Ans. None: 0
Ans. SQuINT: 0
Ans. CP-VQA: 0
Ans. Ours: 0

Ans. None: No
Ans. SQuINT: Yes 🚨
Ans. CP-VQA: No
Ans. Ours: No

Ground Truth

Question 1	Ans. 1	Relation	Question 2	Ans. 2
What is the DME grade for this image?	0	→	Are there hard exudates in this region?	No

Ans. None: 0
Ans. SQuINT: 0
Ans. CP-VQA: 0
Ans. Ours: 0

Ans. None: Yes 🚨
Ans. SQuINT: Yes 🚨
Ans. CP-VQA: No
Ans. Ours: No

Ground Truth

Question 1	Ans. 1	Relation	Question 2	Ans. 2
What is the DME grade for this image?	2	→	Are there hard exudates in the image?	Yes

Ans. None: 2
Ans. SQuINT: 0
Ans. CP-VQA: 0
Ans. Ours: 2

Ans. None: Yes
Ans. SQuINT: No
Ans. CP-VQA: No
Ans. Ours: Yes

Ground Truth

Question 1	Ans. 1	Relation	Question 2	Ans. 2
What is the DME grade for this image?	0	→	Are there hard exudates in this region?	No

Ans. None: 0
Ans. SQuINT: 0
Ans. CP-VQA: 0
Ans. Ours: 0

Ans. None: Yes 🚨
Ans. SQuINT: No
Ans. CP-VQA: No
Ans. Ours: Yes 🚨

Ground Truth

Question 1	Ans. 1	Relation	Question 2	Ans. 2
What is the DME grade for this image?	2	→	Are there hard exudates in the image?	Yes

Ans. None: 0
Ans. SQuINT: 0
Ans. CP-VQA: 0
Ans. Ours: 2

Ans. None: No
Ans. SQuINT: No
Ans. CP-VQA: No
Ans. Ours: Yes

FIGURE D.5: Additional qualitative examples from the DME dataset using MVQA as the backbone. Red siren symbols indicate inconsistent cases. DME is a disease that is staged into grades (0, 1 or 2), which depend on the number of visual pathological features of the retina.

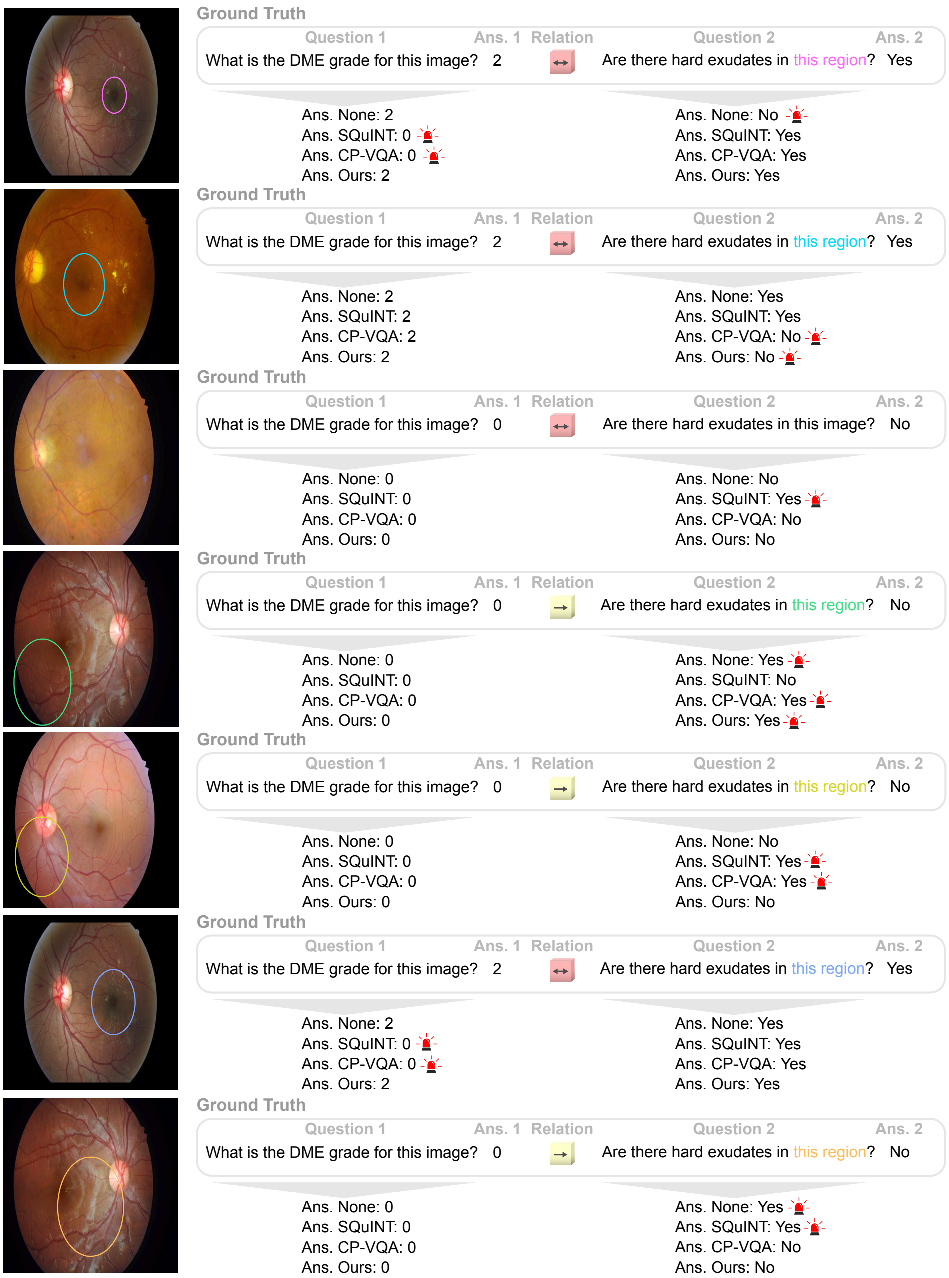

Ground Truth
Question 1 | Ans. 1 | Relation | Question 2 | Ans. 2
What is the DME grade for this image? 2 ↔ Are there hard exudates in this region? Yes
Ans. None: 2
Ans. SQuINT: 0 🚨
Ans. CP-VQA: 0 🚨
Ans. Ours: 2
Ans. None: No 🚨
Ans. SQuINT: Yes
Ans. CP-VQA: Yes
Ans. Ours: Yes
Ground Truth
Question 1 | Ans. 1 | Relation | Question 2 | Ans. 2
What is the DME grade for this image? 2 ↔ Are there hard exudates in this region? Yes
Ans. None: 2
Ans. SQuINT: 2
Ans. CP-VQA: 2
Ans. Ours: 2
Ans. None: Yes
Ans. SQuINT: Yes
Ans. CP-VQA: No 🚨
Ans. Ours: No 🚨
Ground Truth
Question 1 | Ans. 1 | Relation | Question 2 | Ans. 2
What is the DME grade for this image? 0 ↔ Are there hard exudates in this image? No
Ans. None: 0
Ans. SQuINT: 0
Ans. CP-VQA: 0
Ans. Ours: 0
Ans. None: No
Ans. SQuINT: Yes 🚨
Ans. CP-VQA: No
Ans. Ours: No
Ground Truth
Question 1 | Ans. 1 | Relation | Question 2 | Ans. 2
What is the DME grade for this image? 0 → Are there hard exudates in this region? No
Ans. None: 0
Ans. SQuINT: 0
Ans. CP-VQA: 0
Ans. Ours: 0
Ans. None: Yes 🚨
Ans. SQuINT: No
Ans. CP-VQA: Yes 🚨
Ans. Ours: Yes 🚨
Ground Truth
Question 1 | Ans. 1 | Relation | Question 2 | Ans. 2
What is the DME grade for this image? 0 → Are there hard exudates in this region? No
Ans. None: 0
Ans. SQuINT: 0
Ans. CP-VQA: 0
Ans. Ours: 0
Ans. None: No
Ans. SQuINT: Yes 🚨
Ans. CP-VQA: Yes 🚨
Ans. Ours: No
Ground Truth
Question 1 | Ans. 1 | Relation | Question 2 | Ans. 2
What is the DME grade for this image? 2 ↔ Are there hard exudates in this region? Yes
Ans. None: 2
Ans. SQuINT: 0 🚨
Ans. CP-VQA: 0 🚨
Ans. Ours: 2
Ans. None: Yes
Ans. SQuINT: Yes
Ans. CP-VQA: Yes
Ans. Ours: Yes
Ground Truth
Question 1 | Ans. 1 | Relation | Question 2 | Ans. 2
What is the DME grade for this image? 0 → Are there hard exudates in this region? No
Ans. None: 0
Ans. SQuINT: 0
Ans. CP-VQA: 0
Ans. Ours: 0
Ans. None: Yes 🚨
Ans. SQuINT: Yes 🚨
Ans. CP-VQA: No
Ans. Ours: No

FIGURE D.6: Additional qualitative examples from the DME dataset using MVQA as the backbone. Red siren symbols indicate inconsistent cases.

Declaration of Originality

Last name, first name: Tascón Morales Sergio

Matriculation number: 20-140-794

I hereby declare that this thesis represents my original work and that I have used no other sources except as noted by citations.
All data, tables, figures and text citations which have been reproduced from any other source, including the internet, have been explicitly acknowledged as such.
I am aware that in case of non-compliance, the Senate is entitled to withdraw the doctorate degree awarded to me on the basis of the present thesis, in accordance with the "Statut der Universität Bern (Universitätsstatut; UniSt)", Art. 69, of 7 June 2011.

Place, Date: **Bern, March 6, 2024**

Signature:

Sergio Tascón M