Reasoning and Question Answering About Image-Text Multi-modal Contexts

by

Shailaja Keyur Sampat

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved August 2024 by the
Graduate Supervisory Committee:

Chitta Baral, Chair
Yezhou Yang
Nakul Gopalan
Prescott Klassen
Eduardo Blanco

ARIZONA STATE UNIVERSITY

December 2024

ABSTRACT

The ability of autonomous systems to ingest multimedia content and generate a high-level semantic understanding of it is invaluable in many real-world applications. While the advancements in this research area are remarkable and impactful, existing vision-language (V&L) systems are yet to explore many aspects of human cognition. This dissertation addresses two such directions- (1) joint reasoning over multiple modalities and (2) reasoning beyond information perceived from the visual content. In most image understanding tasks, the model is evaluated based on its ability to use language to communicate about the contents of the image (such as classifying or generating text in the form of sentences or short phrases). On the other hand, humans often integrate information from images and accompanying text to make decisions in day-to-day life, for example, product assembly, navigation, and interpreting charts. Towards this end, a novel V&L task is proposed, which requires joint reasoning over images and text. A new multi-task corpus is prepared to support this task, and benchmark models are developed. Further, through a case study on a large-scale synthetic dataset, it is shown that the proposed task formulation can be beneficial to train systems that can reason about images beyond pixel-level information. In particular, the model's ability to perform hypothetical reasoning about actions is investigated (that is, given an image of the world and a description of an imaginary action to be performed, predict the next state of the world). In addition, broader aspects of action reasoning are explored, which includes predicting preconditions, associated high-level goals, effects, and temporal dynamics about actions perceived in cooking images. A few-shot method is proposed to distill knowledge from Large Language Models about the above aspects of actions, conditioned on the visual information. Finally, the contributions of this thesis and its potential real-world applications are summarized, along with a discussion about future scopes in this research direction.

DEDICATION

To my parents (Tsweti & Keyur) and my brother (Krunal)

ACKNOWLEDGMENTS

The completion of this dissertation would not have been possible without the support and encouragement of many individuals. I want to take this opportunity to express my gratitude to all of them.

I want to express my gratitude to my advisor, Dr. Chitta Baral, for his guidance, support, and patience throughout the doctoral program. Working under his supervision helped me develop a strong foundation in my research area and gain independent research skills. His thoughtful input, from choosing the right research problem to developing viable solutions, has been invaluable in shaping my thesis. I am thankful for the positive impact of his support and mentorship as I begin my career as a researcher.

I also want to thank Dr. Yezhou Yang, Dr. Nakul Gopalan, Dr. Prescott Klassen, and Dr. Eduardo Blanco for being a part of my thesis defense committee and providing insightful feedback and suggestions on my work. I am grateful to my collaborators and peers in Dr. Baral's lab and Dr. Yang's group at ASU for their fruitful discussions on a broad range of topics and for challenging me with different perspectives on my research work.

My parents, Keyur Sampat, and Tsweti Sampat are the two persons who are most excited for this day. It is both of you, who showed me the sky, gave me the wings, and let me fly. Thank you for believing in me, and your constant love and support over the years. I am thankful to my fauji brother Krunal Sampat for being there whenever I needed him. Despite our different career pathways, observing your perseverance and composure through ups and downs has helped me to stay focused and disciplined. Last, but not the least, I would like to thank Subhasish Das, whom I met halfway through this journey and he has been my companion and cheerleader since then.

TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

CHAPTER Page

LIST OF TABLES

Table Page

Table Page

Table Page

Table Page

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Chapter 1

INTRODUCTION

1.1 Vision-Language Integration: Artificial Intelligence (AI) Perspective

Images and videos have become an inseparable part of modern communication. The widespread accessibility and affordability of computing devices that can store, process, and display such visual content has resulted in an abundance of availability of visual data. Therefore, it is important to design systems that can ingest such large-scale data and perform high-level semantic understanding and reasoning over it to identify critical events, explain novel concepts, or influence the decision-making process. Development of such systems is crucial for industries including - but not limited to - healthcare, fashion, defense, education, and social media analytics. The related area of research is broadly known as 'Image Understanding'.

Formally, image understanding is a sub-field of Computer Vision (CV) that aims to locate, characterize, and recognize objects and corresponding attributes from the images. Early approaches in image understanding were focused on extracting low-level features (like edges and corners) to match simple visual patterns. Due to factors such as the recent success of deep learning methods, availability of large-scale annotated visual data, and high-performance computing hardware, there have been an unprecedented number of advancements in this field. As a result, the image understanding community has been successful in developing promising solutions to several tasks related to the extraction of mid-level features such as fine-grained image segmentation, object detection, and recognition.

The ultimate goal of developing Artificial Intelligence (AI) agents is to automate

and/or assist humans in performing everyday tasks. For humans, spoken language is considered to be the primary mode of communication. Therefore, there is a growing demand for the development of interfaces that can facilitate humans and machines to communicate effectively. While visual modalities provide a concise representation of the physical world, language makes it easier to express concepts that connect the physical world with other kinds of knowledge. Consequently, a wide variety of AI research prototypes have been developed that incorporate the integration of vision and language. This includes image understanding models guided by linguistic cues, and multi-modal conversational agents to robots that perform tasks instructed in natural language.

In recent years, Natural Language Processing (NLP) has evolved separately as a sub-branch of AI. However, research progress in CV and NLP has had a great influence on each other due to the nature of the problems addressed in these research fields. In the NLP domain, the task of Question Answering (QA) is well accepted for evaluating the level of understanding of a system. It is largely inspired by the educational setting, where a student's understanding of a subject is measured by their ability to solve problems of varying levels of difficulty.

The image understanding community has widely adopted the QA style of evaluation used in NLP and is popularly known as the Visual Question Answering (VQA) task. The underlying hypothesis for the VQA task states that if an AI system has a semantic understanding of a visual scene, it should be able to answer natural language questions around those semantics. In other words, if an AI system is provided with an image and posed with several questions of varying difficulty related to the given image, then the system's understanding of the visual content can be measured by its ability to answer given questions correctly.

In recent years, a large body of visual question answering (VQA) benchmarks have

been compiled to facilitate the development of AI systems and to assess progress in this field of research. This introductory chapter provides an overview of existing datasets that have been instrumental in advancing VQA research. Following this, several open challenges and unexplored research directions in image understanding are discussed that are inspired by human cognitive abilities. This will provide the foundation for the remainder of this dissertation.

1.2 Background: Visual Question Answering (VQA)

The aim of the Visual Question Answering (VQA) task is to facilitate the development of systems that can answer natural language questions about the contents of visual information. In most cases, a set of possible answers that are compiled or known beforehand allows the formulation of VQA as a classification problem. Formally, the VQA task consists of three components: an image I, a natural language question Q, and a set of possible answer choices A. The goal is to develop a model f_θ (parameterized by θ) that can select answer $\mathtt{a}$ from a list of candidate answers A for a given problem (I, Q), i.e.,

$$f_\theta(I, Q) \rightarrow \mathtt{a} \in A \tag{1.1}$$

One can ask a wide variety of questions- ranging from a particular visual attribute of an entity and interrelation of various objects to a high-level context conveyed by the image. Therefore, VQA task formulation allows models to comprehend visual information at both the global and local level (Mogadala *et al.*, 2021). As a result, VQA has been one of the widely used task formats in the vision-language domain. As per a survey, 53 among 130 benchmarks for visual understanding (as of 2020) supported VQA task formulation. While the earlier approaches (Geman *et al.*, 2015; Malinowski and Fritz, 2014) considered VQA as a Visual Turing Test and attempted

to answer simple binary questions, the current efforts are focused towards large-scale, open-ended and application-specific VQA datasets (Antol *et al.*, 2015b; Marino *et al.*, 2019; Kahou *et al.*, 2017; Ben Abacha *et al.*, 2019). This section discusses important existing VQA datasets. They are organized into three major clusters depending on whether questions are answerable from images only, would require any knowledge beyond the image, or require reasoning over more than one modality. While visuals typically include both images as well as videos, this survey is limited to benchmarks that involve images.

1.2.1 VQA for Information Explicitly Present in Images

In this cluster of the VQA datasets, questions are centered around information explicitly present in the image which includes objects, their attributes, or ongoing actions and events (as shown in Figure 1.1). DAtaset for QUestion Answering on Real-world images (DAQUAR) proposed by Malinowski and Fritz (2014) was one of the earliest publicly available VQA datasets. It incorporated three kinds of questions-object type, object color, and number of objects over indoor images annotated with segments, object labels, and depth values. Later, the COCO-QA dataset (Ren *et al.*, 2015a) was proposed where authors automatically converted an image-captioning dataset MS-COCO (Lin *et al.*, 2014) into a VQA. A parallel line of work by Antol *et al.* (2015a) leveraged crowd-sourcing for collecting question-answer pairs over MS-COCO images.

Following the success of the above pioneers, several variants of visual QA have been proposed. Among them, VQAv2 (Goyal *et al.*, 2017), Visual7w (Zhu *et al.*, 2016), and VizWiz (Gurari *et al.*, 2019) focused on questions about objects, their attributes, actions, and relationships among objects. Datasets such as CountQA (Chattopadhyay *et al.*, 2017), HowManyQA (Trott *et al.*, 2018) and TallyQA (Acharya *et al.*, 2019)

Figure 1.1: VQA Examples From Johnson *et al.* (2017a); Ren *et al.* (2015a); Hudson and Manning (2019) Where Information Contained in the Image is Sufficient to Answer a Given Question

are specifically designed for questions involving complex object counting. On the other hand, datasets like NLVR (Suhr *et al.*, 2017) and GQA (Hudson and Manning, 2019) were developed to improve the spatial reasoning capability of systems through question answering. The TextVQA (Singh *et al.*, 2019), OCR-VQA (Mishra *et al.*, 2019), and ST-VQA (Biten *et al.*, 2019) benchmarks were developed with the goal of understanding text inside images, which might be relevant when images are taken in the marketplace, restaurants, streets, etc.

All the above datasets were based on natural images (i.e. everyday indoor/outdoor scenes that humans encounter frequently) and leveraged a combination of automated methods and crowd-sourcing for large-scale data collection. While natural images-based VQA datasets reflect challenges one can encounter in real-life situations, the requirement of costlier human annotations and vulnerability to biases are two major drawbacks. To overcome this, synthetic dataset generation approaches were proposed. Synthetic approach allows controlled data generation at scale while being flexible to test specific reasoning skills. As a result, several synthetic 2D and 3D VQA datasets have been developed which include COG (Yang *et al.*, 2018), Shapes (Andreas *et al.*,

2016), CLEVR (Johnson *et al.*, 2017a) and CLEVR-Dialog (Kottur *et al.*, 2019).

The FigureQA (Kahou *et al.*, 2017) and DVQA (Kafle *et al.*, 2018) datasets took synthetic data generation one step further. Specifically, they generated synthetic charts (bar chart, pie chart, dot-line, etc.) and programmatically curated question-answer pairs around them. In between natural and synthetic images lies VQA-abstract (Antol *et al.*, 2015b) and IQA (Gordon *et al.*, 2018) datasets which involve clipart-style scenes and rendered interactive environments respectively.

1.2.2 VQA Requiring External Knowledge

Humans possess a rich knowledge about the surrounding world and utilize that knowledge while they attempt visual reasoning. A similar situation might occur for AI systems where the provided visual information alone might not be sufficient in order to answer many questions, as shown in Figure 1.2. Motivated by this aspect of human reasoning abilities, knowledge-based VQA benchmarks are being developed. The KB-VQA dataset (Wang *et al.*, 2017a) was the first knowledge-enhanced VQA benchmark. Particularly, the dataset consisted of several kinds of templates-generated questions over MS-COCO images (Lin *et al.*, 2014), answering which would require knowledge about the objects (present in images) which is available in the form of knowledge-triples in DBpedia (Auer *et al.*, 2007).

Post that, the F-VQA dataset (Wang *et al.*, 2017b) was proposed as an extension of KB-VQA (Wang *et al.*, 2017a), which was larger and more complex in terms of knowledge. F-VQA leveraged three existing knowledge bases- DBpedia (Auer *et al.*, 2007), ConceptNet (Speer *et al.*, 2017) and WebChild (Tandon *et al.*, 2014) for curation of their data. Another knowledge-intensive task KVQA was proposed by Shah *et al.* (2019), which focused on the world knowledge about named entities in the image (e.g. Barack Obama, White House, United Nations). It is one of the largest

Figure 1.2: VQA Examples From Wang *et al.* (2017a,b); Marino *et al.* (2019) Where Some Additional Common Sense or World Knowledge is Required Beyond the Information Contained in the Image in Order to Answer a Given Question

knowledge-VQA benchmarks to date, which requires multi-entity, multi-relation, and multi-hop reasoning over Wikidata knowledge graph (Vrandečić and Krötzsch, 2014).

The reliance on one or more knowledge graph(s) has two disadvantages; First, querying across knowledge bases requires standardization in terms of storage/format and is only viable under the availability of a query language that works with all knowledge bases. Second, with a higher number of knowledge bases involved, the inference becomes more expensive. On the other hand, humans utilize search engine-like tools to obtain missing knowledge by creating appropriate queries over unstructured documents. Motivated by this, OK-VQA (Marino *et al.*, 2019) was developed, where models solving this dataset require the capability to extract open-ended knowledge from the web in order to answer image-based questions.

This cluster of the VQA datasets also includes resources where domain-specific knowledge might be needed to answer questions. For example, datasets proposed by Ben Abacha *et al.* (2019); Lau *et al.* (2018); Naseem *et al.* (2023) contain medical/-clinical imagery (such as x-rays, radiographs, pathology specimens, etc.) and ask relevant questions involving observations present in the imagery, determining the ab-

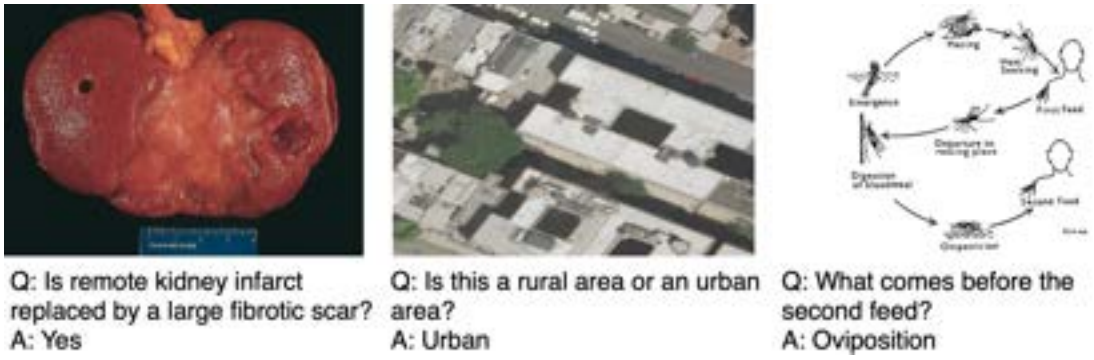


Figure 1.3: VQA Examples From Naseem *et al.* (2023); Lobry *et al.* (2020); Kembhavi *et al.* (2016) Where Domain-specific Knowledge (Such as Mathematics, Science, Medical/Clinical or Remote Sensing Related) is Required Beyond the Information Contained in the Image in Order to Answer a Given Question

sence/presence of particular concepts or predicting underlying conditions of a patient. On the other hand, RS-VQA (Lobry *et al.*, 2020) is built upon satellite imagery and a broad range of QA pairs concerning various remote sensing applications such as classifying terrain types, detecting and counting objects of interest, and monitoring changes over time. VQA datasets based on Mathematics and Science (Seo *et al.*, 2014; Kembhavi *et al.*, 2016) are relevant here as well. Figure 1.3 illustrates representative VQA datasets that require specialized domain knowledge.

1.2.3 Multi-Modal VQA

Multi-modal learning in AI is referred to as building models that can process and relate information from two or more modalities. While multi-modal contexts can broadly include any combinations of audio, text, images, video, etc., image+text multi-modality is most relevant to the VQA task. There is only one dataset TQA (Kembhavi *et al.*, 2017) that is relevant under this category. This dataset is curated from middle school science curricula and provides long lessons (∼50 sentences and

4-5 images) as a context for the AI systems. The goal of TQA is to equip models with the ability to perform a careful selection of necessary facts from the long-tailed contexts in order to answer a question (as demonstrated by examples in Figure 1.4). Although they claim that their task requires multi-modal machine comprehension, 40% of their text-based questions can be answered using a single sentence, and 50% of their image QA can be answered using only one image from the context. In that case, a significant portion of this dataset becomes analogous to machine comprehension or ordinary VQA tasks, losing out on the actual purpose of multi-modality.

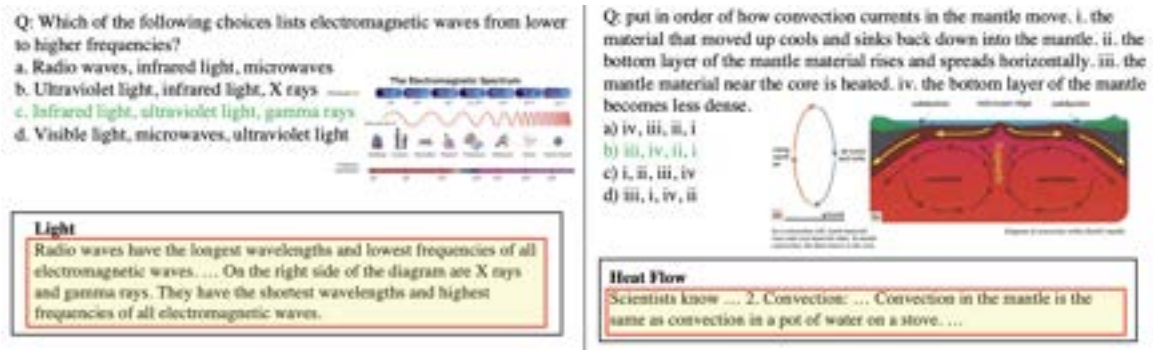


Figure 1.4: Multi-modal Machine Comprehension Task Proposed by Kembhavi *et al.* (2017) With Two Examples From Their TQA Dataset; Each Example in This Dataset Contains Long Visual-linguistic Context (∼50 Sentences and 4-5 Images) in Order to Answer a Given Question. In This Figure, Only a Relevant Portion of the Long Context is Shown (Yellow Box at the Bottom of Each Example)

1.3 Contributions

Despite the significant progress and active efforts in VQA research as discussed above, there are many aspects of human cognition yet to be explored and a considerable number of open challenges exist that need to be tackled. This dissertation explores two such research directions with the long-term goal of advancing current

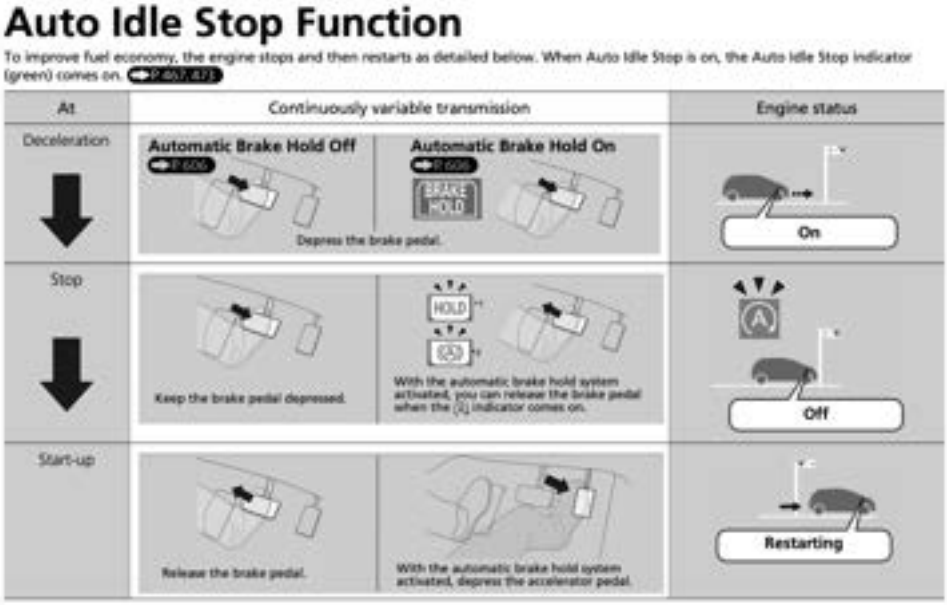
AI systems to human-levels of reasoning abilities and visual content understanding. In particular, this thesis explores the development of AI systems that can perform inference over multiple modalities and reason beyond visually perceivable information (particularly about actions). A brief description motivating the importance of the above two directions from a vision-language applications perspective is provided below. Comprehensive details about both research directions, proposed approaches, and developed solutions are covered in subsequent chapters, as per the outline in Section 1.4.

Multi-modal Understanding

Humans often integrate information from multiple sources (images+text being the most frequent case) to make decisions in day-to-day life. For example, product assembly using instruction manuals, navigating roads while following street signs, interpreting visual representations in various documents such as newspapers and reports, understanding concepts using textbook-style learning, etc. Figure 1.5 visually demonstrates a couple of such use cases. The first scenario demonstrates the use of visual and textual information together to explain the complex functionality of a car. Whereas the second use case depicts two newspaper articles involving a body of text along with charts/infographics that provide additional context about the topic (such as historical data or relative comparison).

The importance of joint reasoning over images and text has also been emphasized in the design of standardized/psychometric tests like PISA and GRE. Programme for International Student Assessment (PISA) tests conducted post-2018 take into account "the evolving nature of reading in digital societies- which requires an ability to compare, contrast and integrate information from multiple sources" (OECD, 2019). Graduate Record Examinations (GRE) has data interpretation questions that assess

(a)



(b)

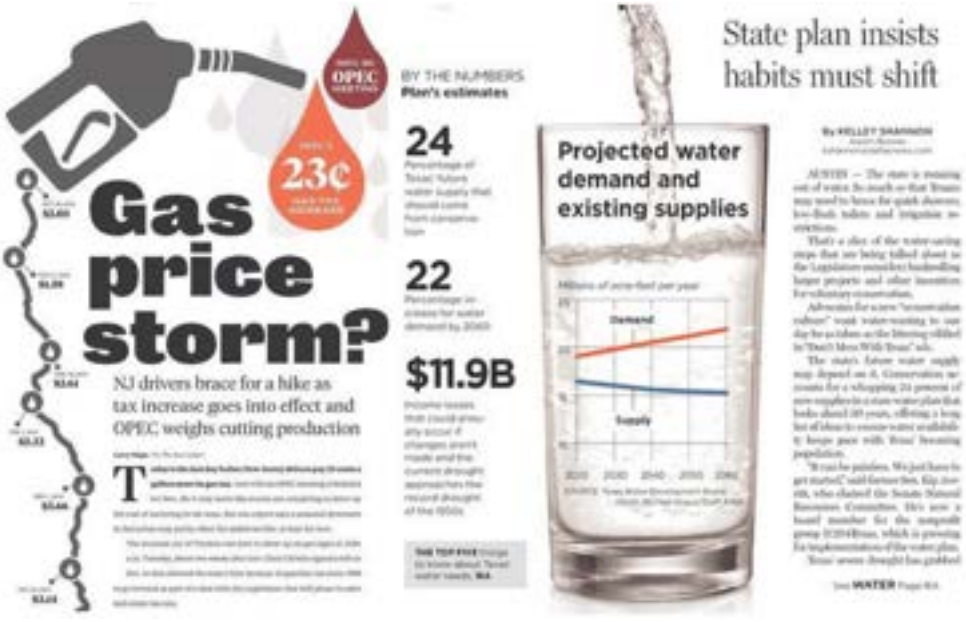


Figure 1.5: Examples Demonstrating the Role of 'Multi-modal Understanding' in Day-to-day Life, Which are More Complex Than Current VQA Benchmarks: (a) A Snippet from Honda User Manual (Honda, 2022) and, (b) Two Pieces of News Articles (Higgs, 2016; Shannon, 2013), Where Visuals and Text Complement Each Other to Aid the Reader With a Better Understanding of the Topic

a student's ability to analyze given data as a combination of text and charts.

In the formulation of image understanding as a VQA task, the system is provided with an image and is expected to correctly answer a broad range of questions related to the image. Therefore, the role of language, in this case, is merely as an evaluation mechanism rather than as a modality that provides additional context about the problem/topic. Both the aforementioned evidence suggest that multi-modal reasoning is crucial for humans in day-to-day life, which is not addressed in VQA research.

To fill in the gap between existing VQA systems and human-like reasoning, this dissertation proposes the Visuo-Linguistic QA (VLQA) task. VLQA is defined as a variant of the VQA task where deriving joint inference between visual and textual information is a necessity to correctly answer posed questions. In other words, provided image(s) and a text passage, the task is to answer questions by joint reasoning over them i.e., ignoring a cue from either modality would make the question unanswerable. To this extent, a new VLQA dataset is compiled and experiments are conducted with state-of-the-art VQA models to gauge their joint reasoning performance using the given image(s) and text. In subsequent work, a large-scale, multi-task visual-linguistic benchmark (VL-GLUE) is developed to encourage the development of vision-language models that have improved multi-modal reasoning capabilities.

Reasoning about Actions and Change

In most existing VQA datasets, scene understanding is holistic and questions are centered around information explicitly present in the image (i.e. objects, attributes, actions). As a result, advanced object detection and scene graph techniques have been quite successful in achieving good performance over existing benchmarks. However, given an image, humans can speculate a wide range of implicit information (i.e. not directly perceivable from the image). Examples include- determining the purpose

of various objects in a scene, how one object relates to other objects, speculating about events that might have happened in the past, considering numerous imaginary situations and predicting possible future outcomes, recognizing the intentions of a subject to perform particular actions, and so on.

The VQA datasets that require external knowledge (discussed in Section 1.2.2) have attempted to address implicit reasoning to some extent. However, a majority of their efforts have been limited to the knowledge that is tightly coupled with given visual entities (e.g. intention of a person exercising is to stay fit or keep the body in shape, Barack Obama is a former US President). There are no existing benchmarks that tackle hypothetical reasoning i.e. a model is posed with a certain imaginary situation and it has to reason about possible outcomes of that situation. This is another important characteristic of human cognition that is unexplored in intelligent agents.

Hypothetical reasoning ability plays a critical role in how humans tackle unanticipated events or novel situations. For example, if any person is asked whether they should 'jump into an active volcano'. Even though one might not have encountered such a situation anytime (i.e. being in the middle of the lava physically), one can perform hypothetical reasoning using the prior and present knowledge related to this scenario: lava is hot, hot things burn, burns are painful, painful things are bad. Using such a reasoning chain, one can conclude that it is not a good idea.

A similar ability (to perform hypothetical reasoning) would be highly desirable in autonomous agents which assist humans in performing everyday tasks. Figure 1.6 demonstrates this aspect through an example. Consider a robot that is instructed to 'pour wine into a glass'. The simplest case is shown in Scenario (a), where a robot grabs a wine bottle and pours the wine into a wine glass. However, sometimes different situations might arise. For example, the robot finds the glass broken where

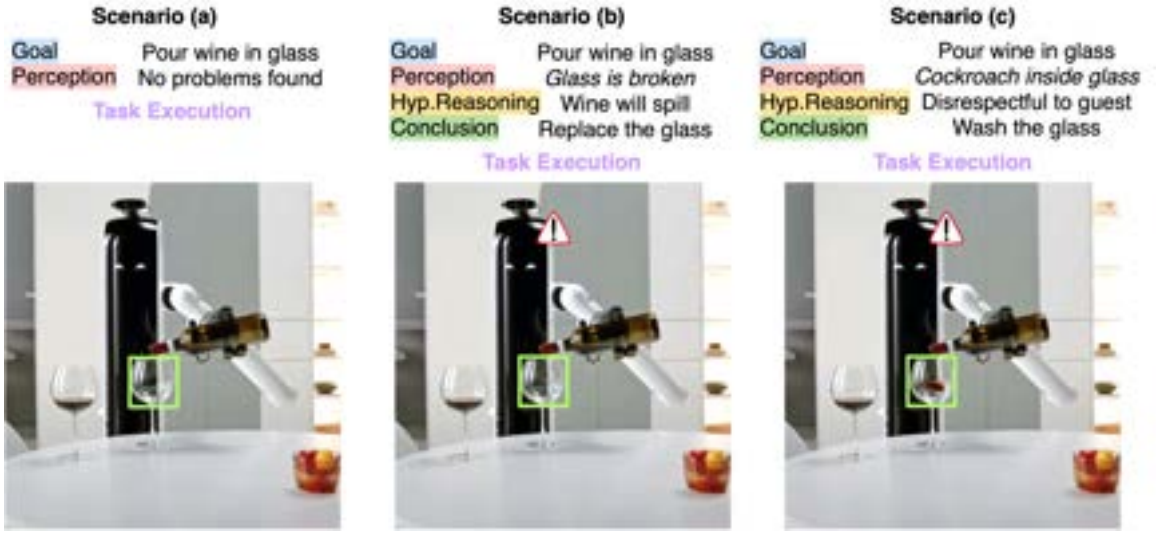


Figure 1.6: Example Demonstrating the Importance of Hypothetical Reasoning Ability in AI Agents That Will Perform Day-to-day Tasks: Three Scenarios Encountered While a Robot Attempts to 'Pour Wine Into a Glass' (a) No Unanticipated Event Identified, (b) Robot Finds the Glass Broken, and (c) Robot Finds a Cockroach Inside the Glass

the wine is to be poured, as shown in Scenario (b). In this case, the robot requires to perform hypothetical reasoning that 'pouring wine into a broken glass will cause the wine to spill over the table'. This will be helpful to decide the next course of action i.e. 'replace the broken wine glass with a good one and then pour wine into it'. Similarly, the robot might find a cockroach inside the glass i.e. Scenario (c). In this case, hypothetical reasoning 'serving wine with a cockroach inside will be disrespectful to the guest' is helpful to decide the next action 'wash the glass before pouring wine'.

Motivated by the aforementioned use cases, this dissertation explores to what extent intelligent agents can be trained to do hypothetical reasoning. A particular focus of this work is on hypothetical reasoning about actions due to its usefulness in many vision-language applications. For this study, a new synthetic VQA benchmark

CLEVR_HYP (Hypothetical Reasoning about CLEVR-style images) is developed. Specifically, a visual scene with rendered objects is provided and the model has to perform an imaginary action over the scene that is described in the natural language. The model's capability to perform hypothetical reasoning is evaluated by its ability to answer questions about the resulting scene post-action execution. Several existing neural and neural-symbolic architectures' capability on this hypothetical reasoning task is benchmarked. Further, a novel three-stage action representation learning strategy is proposed to better capture the causal structure of this task.

As the CLEVR_HYP benchmark is synthetically rendered, it considers a limited number of possible actions and object attributes. The benchmark also assumes that actions are always executable and guaranteed to produce visual effects. This might not be the case in real-world scenarios. There are several other factors associated with actions besides effects, which require reasoning beyond visual information. These factors include preconditions that must be fulfilled for the execution of actions (e.g. only liquid objects can undergo pouring), reasoning about high-level goals associated with actions (e.g. beat the eggs to make an omelet), and temporal dynamics among actions (e.g. crack eggs before whisking or draining pasta after boiling).

Considering this complex nature of action reasoning, a more realistic benchmark is curated from an annotated cooking video dataset incorporating the aforementioned aspects of actions. In particular, provided an image where some actions are ongoing, have recently finished or are about to take place, the task is to determine preconditions, possible effects, associated high-level goals, and likely actions that may precede or follow. The recent developments in Large Language Models (LLMs) indicate that such models possess a vast amount of information about the world due to training on massive amounts of data. Experiments are conducted to assess how accurately different aspects of actions are captured by LLMs (specific to the provided visual

input). Ablation studies are used to investigate how to effectively prompt LLMs to discern such knowledge.

1.4 Organization of the Dissertation

In this chapter, existing efforts related to Visual Question Answering (VQA) are concisely summarized. Based on this survey, two research directions are identified, which are motivated by the gaps in current benchmarks and their potential applications. The remainder of this dissertation prospectus is organized as follows; In Chapter 2, a novel multi-modal VQA task and dataset is proposed, where combining clues from the given visual and textual contexts are necessary to answer questions. This chapter covers a detailed description of the dataset curation process, analysis of the dataset from different aspects and baselines developed using off-the-shelf architectures. Chapter 3 investigates the hypothetical reasoning abilities of AI systems with a specific focus on the effects of the actions. This is considered a special case of the multi-modal VQA task proposed in Chapter 2, where textual contexts describe actions that need to be performed over the objects present in the image. Chapter 4 describes a novel three-stage representation learning strategy that better captures the causal structure of the hypothetical action reasoning task. Additionally, the effectiveness of the proposed three-stage method in terms of data efficiency and generalization capability is demonstrated.

There has been growing interest in multi-modal VQA research post the proposal of the VLQA dataset introduced in Chapter 2. Since then, several other multi-modal VQA datasets have been developed that span a variety of visuals, and test numerous reasoning abilities and in terms of knowledge required to answer questions. In Chapter 5, a multi-task benchmark for visual-linguistic reasoning is developed by bringing together different datasets in a common format and organizing them into

seven clusters. In Chapter 6, Large Language Models (LLMs) are leveraged to discern knowledge about broader aspects of action reasoning (including preconditions, goals, effects, and likely past/future states). Finally, in Chapter 8, the contributions of this thesis and its potential real-world applications are summarized. The findings from this research are consolidated as a roadmap to drive the future development of vision–language models that are equipped with complex reasoning abilities and able to communicate more effectively with humans.

Chapter 2

VISUO-LINGUISTIC QUESTION ANSWERING (VLQA): A NEW MULTI-MODAL VQA BENCHMARK

2.1 Motivation

Question answering (QA) is a crucial way to evaluate an Artificial Intelligence (AI) system's ability to understand text and images. In recent years, a large body of natural language QA (NLQA) datasets and visual QA (VQA) datasets have been compiled to evaluate the ability of a system to understand text and images. For most VQA datasets, the text is used merely as a question-answering mechanism rather than an actual modality that provides contextual information. On the other hand, deriving inference from combined visual and textual information is an important skill for humans to perform day-to-day tasks. For example, product assembly using instruction manuals, navigating roads while following street signs, interpreting visual representations (e.g., charts) in various documents such as newspapers and reports, understanding concepts using textbook-style learning, etc.

The importance of joint reasoning has also been emphasized in the design of standardized/psychometric tests like PISA and GRE[1], as evident from Figure 2.1. PISA assessments conducted post-2018 take into account "the evolving nature of reading in digital societies- which requires an ability to compare, contrast and integrate information from multiple sources" (OECD, 2019). The GRE has data interpretation questions that assess a student's ability to analyze given data as a combination of text and charts.

[1] `https://www.oecd.org/pisa/`, `https://www.ets.org/gre/`

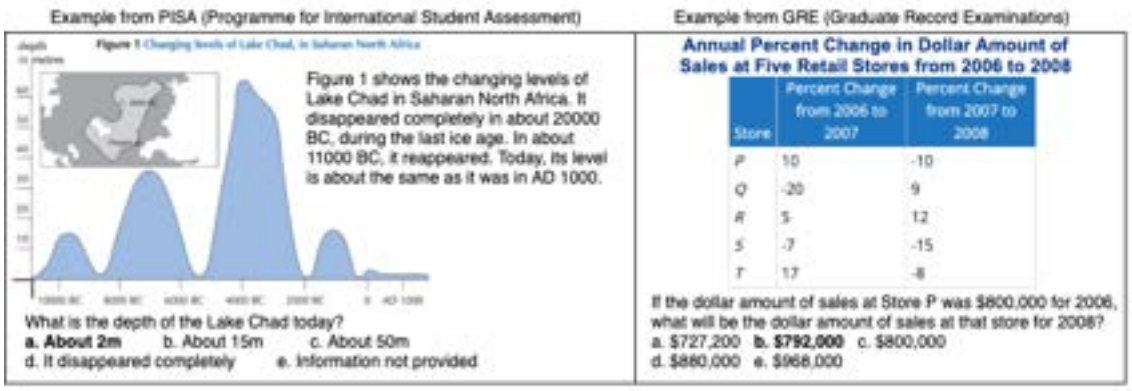


Figure 2.1: Examples of Joint Reasoning Questions in Standardized Tests (Boldface Represents Correct Answer)

Both the aforementioned evidence motivates the need to develop Visuo-Linguistic QA (VLQA) systems, posing a further challenge to state-of-the-art VQA research. As discussed in Sections 1.2 and 1.3, existing VQA research lacks a multi-modal understanding benchmark. In this work, the task of deriving joint inference over images and text is formalized. It is a variant of a VQA where deriving joint inference between visual and textual information is a necessity to correctly answer the question. A dataset is compiled to benchmark the performance of existing VQA models on the proposed VLQA task.

An example of a VLQA task is shown in Figure 2.2. Specifically, there are three components- an image, a text passage, and a multiple-choice question. In the example, the question asks 'How many common followers do the pop singers have?'. To answer this question, one has to understand that 'Tim and Justin are pop singers' (from the text passage) and they have two common followers Maria and Joan (based on the image). Thus, answering this question requires a joint understanding of both the image and text passage. In other words, ignoring a cue from either modality would leave the question unanswerable.

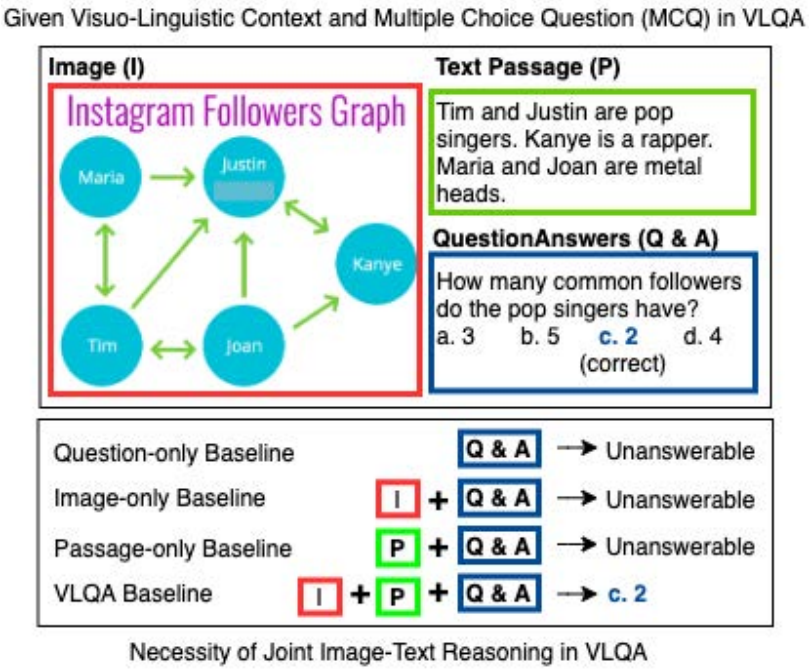


Figure 2.2: Example of Visuo-Linguistic Question Answering (VLQA) Task for Joint Reasoning Over Image-Text Context

2.2 VLQA Task Overview

All data points in the VLQA task format are 5-tuple <Image(I), Text Passage (P), Question (Q), Answer Choices (A), Label (L)>. Each component is explained below in detail.

Image(I): It is provided imagery, which ranges from daily life scenes, and a variety of data representations (e.g. charts) to complex diagrams. A portion of VLQA examples also requires reasoning over multiple images. For the simplicity of processing and retrieval, all images are composed into a single file. Each image is bounded by a red box and provided an explicit detection tag ([0],[1],..) for identification purposes, inspired by VCR dataset (Zellers *et al.*, 2019). This provides a convenient way to reference images in passages, questions, or answers such as Image [0], Image [1], and

so on.

Passage(P): It is a textual modality that provides additional contextual information related to the image. The passages in the VLQA dataset are composed of 1-5 sentences, which consist of facts, imaginary scenarios, or their combination.

Question(Q): It is a question in natural language that tests the reasoning capability of a model over a given image-passage context. In addition to standard 'Wh' patterns and fact-checking style of evaluation (True/False), some questions in VLQA are of 'do-as-directed' form, similar to standardized tests.

Answer Choices(A) and Label (L): VLQA is formed as a classification task over 2-way or 4-way plausible choices, with exactly one of the candidate answers being correct. Answer choices may contain boolean, alpha-numeric phrases, image tags, or their combination.

Task: As described above, each VLQA dataset item is a collection of 5-tuple <I, P, Q, A, L>. The task is to build an AI model f_θ (parameterized by θ) that can select an answer $\mathtt{a}$ (from the set of candidate answers A) for a given question (Q) using image-text multi-modal context (I+P). The correctness of the prediction is measured against the ground-truth label (L). Formally, the VLQA task can be described as follows:

$$f_\theta(I, P, Q) \rightarrow \mathtt{a} \in A, \mathtt{a} = L \tag{2.1}$$

More examples are provided in Figure 2.3. Additionally, each example is provided with rich annotations and classification on several aspects such as image types, question types, required reasoning capability, and the need for external knowledge.

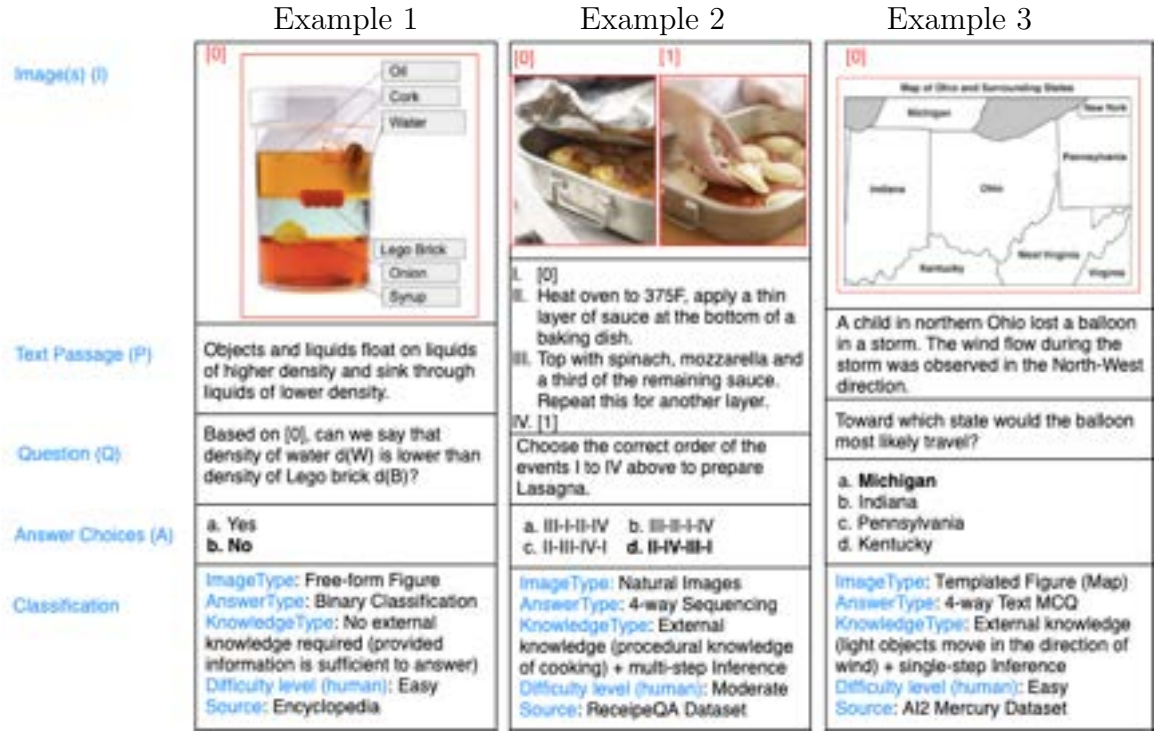


Figure 2.3: Examples From VLQA Training Set: Each Example Contains an Image, a Corresponding Text Passage, and a Multiple Choice Question With the Ground-truth Label Highlighted by the Boldface. Further, Each Sample Contains Annotations Describing Image Type, Answer Type, Requirement of External Knowledge, Reasoning Type, and Human Annotated Difficulty Level

Though the use of this metadata during training is optional, it is useful if one would like to tackle specific subsets of VLQA or in the analysis of a proposed system on various dimensions. A detailed description of the annotations is provided in Section 2.3. Note that, often, the additional text and question are combined in standardized tests, but here they are segregated into Passage and Question for the ease of processing and structured dataset design. The creation of the VLQA dataset is purely research-oriented. By referring to standardized tests as an inspiration, comparison with professional organizations like ETS or OECD is not intended.

2.3 VLQA Dataset Creation

As discussed in Chapter 1.2, there are no VQA benchmarks available for multi-modal understanding. Thus, the main goal of this work is to collect a dataset that requires deriving joint inference from image-text modality. Figure 2.4 illustrates the complete dataset creation process, which has three key stages- Data Collection, Annotation, and Quality Control.

2.3.1 Data Collection

VLQA dataset consists of text together with a diverse range of visual elements. Since manuals, documents, and books containing texts and visuals are ubiquitous, the VLQA dataset is very much grounded in the real world. The dataset is curated from multiple resources (such as books, encyclopedias, web crawls, existing datasets, etc.) through combined automated and manual efforts. The resulting dataset consists of 9267 image-passage-QA tuples with detailed annotation, which are meticulously crafted to assure its quality.

The data sources are segregated into Primary and Secondary; The raw textual/visual information that can be obtained from various web resources, are referred to as primary sources. For example, text crawls from Wikipedia containing facts or images crawled by keyword search. Similarly, tabular data from CIA 'world factbook' (Central Intelligence Agency, 2019) and WikiTables (Pasupat and Liang, 2015) are collected. Later, they are used as it is or post-processed according to the nature of the data. For example, tabular data can be converted to various figures like bar charts, pie charts, scatter plots, etc. Once pre-processed, multiple choice questions are formulated around them. Figure 2.5(a) demonstrates an example that was curated around a figure obtained from a Wikipedia page (as a primary resource).

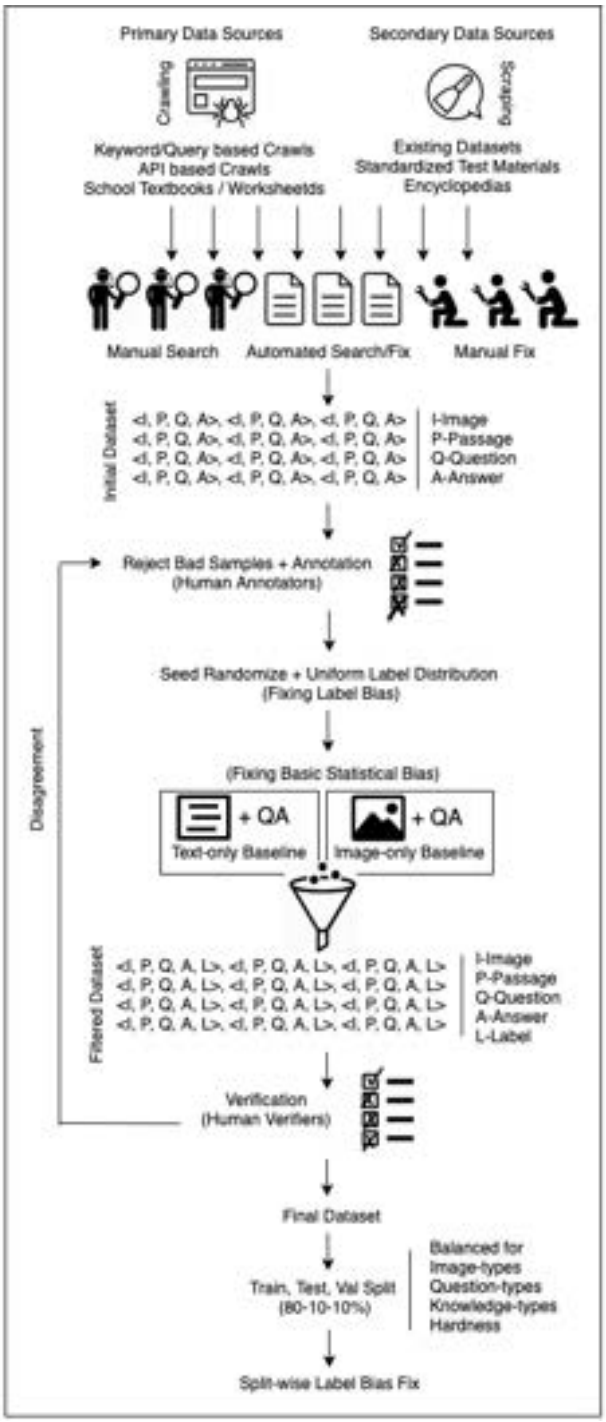
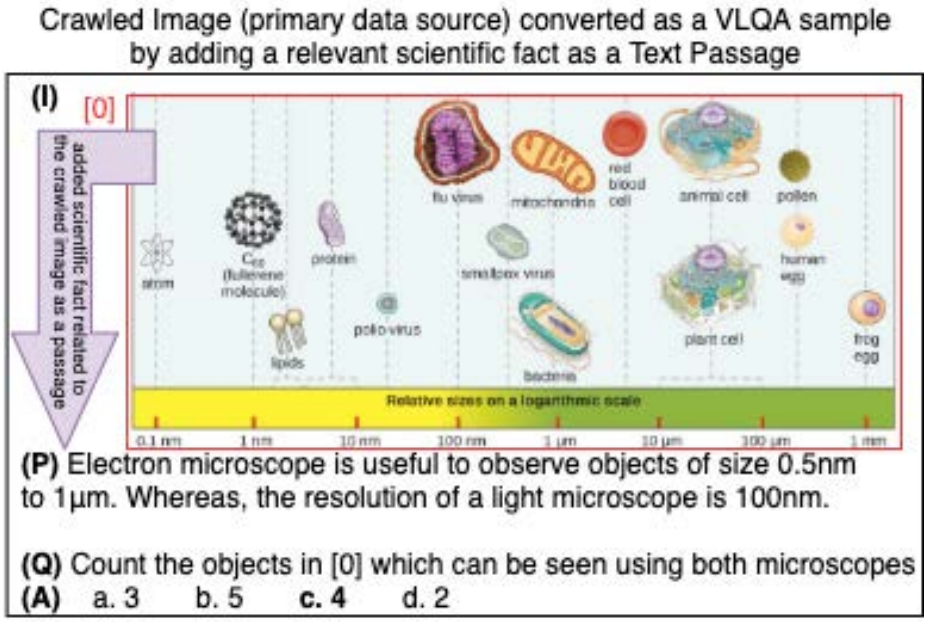
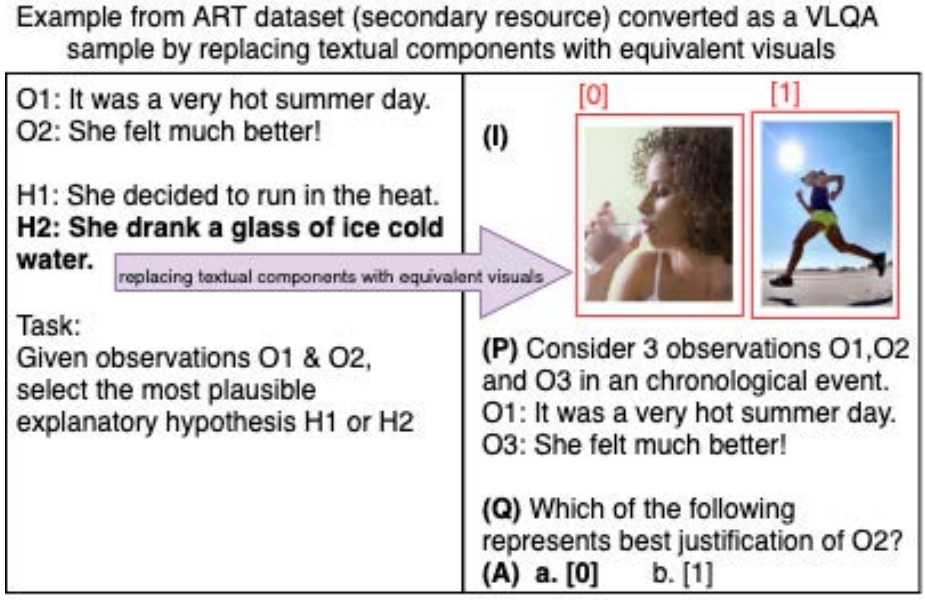


Figure 2.4: Three-stage VLQA Dataset Preparation Which Includes Data Collection From Primary and Secondary Sources, Processing Them Into the Desired Format, and Integrity Steps Performed to Ensure the Quality of the Samples



(a)



(b)

Figure 2.5: Stage-1 of VLQA Dataset Preparation: Data Collection Using Primary and Secondary Sources Followed by Post-processing (if Any), and Finally Creating Question-Answers That Require Joint Reasoning

Existing structured or semi-structured materials are considered secondary data sources, which can be quickly manipulated to use for this dataset; educational materials, standardized tests, and existing vision-language datasets are important from this aspect. Scrapers are used to collect textbook exercises, encyclopedias, practice worksheets, and question banks. Further, a subset of interesting samples from existing datasets is collected- such as RecipeQA (Yagcioglu *et al.*, 2018), wikiHow (Koupaee and Wang, 2018), PhysicalIQA (Bisk *et al.*, 2020a), ART (Bhagavatula *et al.*, 2019) and TQA (Kembhavi *et al.*, 2017).

The collected textual/visual information from the aforementioned sources is refactored and molded as per the VLQA task requirements. Figure 2.5(b) illustrates the data creation process using ART dataset (Bhagavatula *et al.*, 2019) as a secondary resource. Refactoring includes manual or semi-automated post-processing such as replacing given textual/visual attributes with equivalent visual/textual counterparts, adding/removing partial information to/from text or visuals, and creating factual or hypothetical situations around images. By repeating this process over multiple primary/secondary resources and standardizing all the information into question-answers, the initial version of the dataset is obtained.

2.3.2 Human Annotation

Since the VLQA task imposes the condition that a question must be answered through joint reasoning over both modalities, the annotation process becomes non-trivial and requires careful manual annotation. A limited number of in-house expert annotators are preferred for quality purposes rather than a noisier hard-to-control crowd-sourcing alternative.

A crowd worker interface is designed using Python-Flask[2] and deployed on a

[2]https://flask.palletsprojects.com/en/1.1.x/

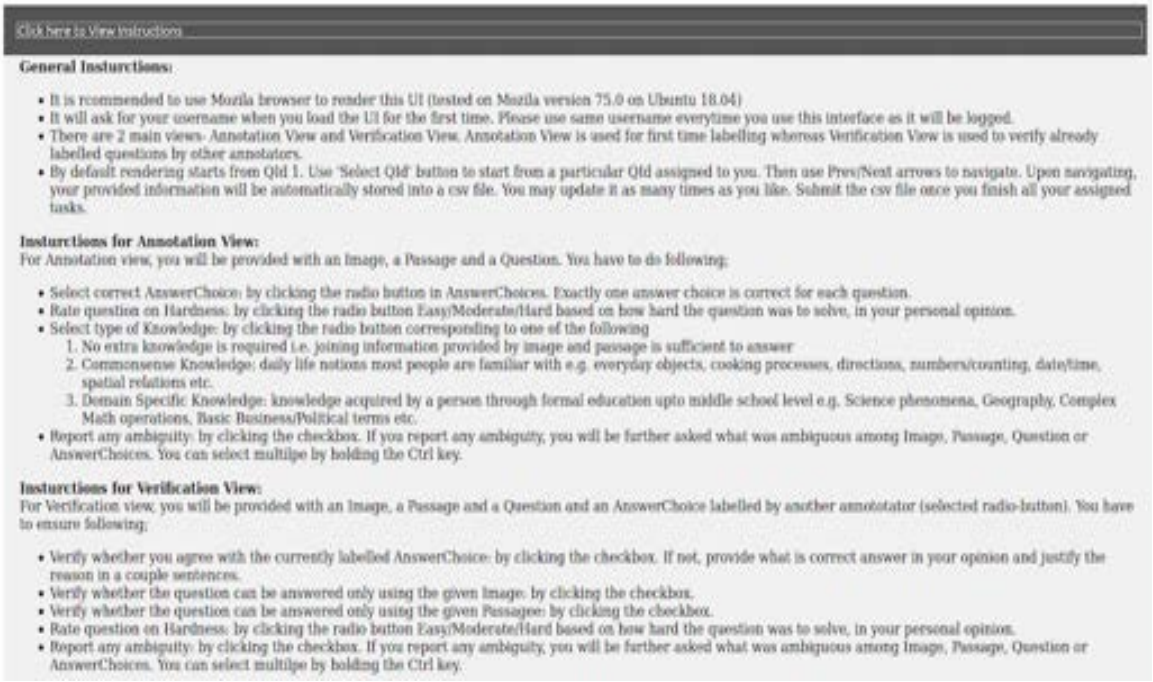


Figure 2.6: Three-fold Instructions Provided to Annotators for VLQA Data Labeling

local server. The data entered by annotators are then logged into Comma Separated Value (CSV) files in a structured format. Annotators are clearly instructed (as per Figure 2.6) about the annotation procedure and it was known to them that exactly one answer choice is correct for each item in the dataset. During the first round of annotations, annotators were allowed to reject bad samples based on two criteria; first, the image and passage must not contain entirely overlapping information and second, a question must not be answerable without looking at the image and passage. For the second criterion, they were instructed to mark such samples as 'ambiguous' as per Figure 2.7.

2.3.3 Ensuring Dataset Integrity

A combined understanding of visual and textual inputs is a key aspect of the VLQA task. As it is modeled as a classification task, some systems might exploit

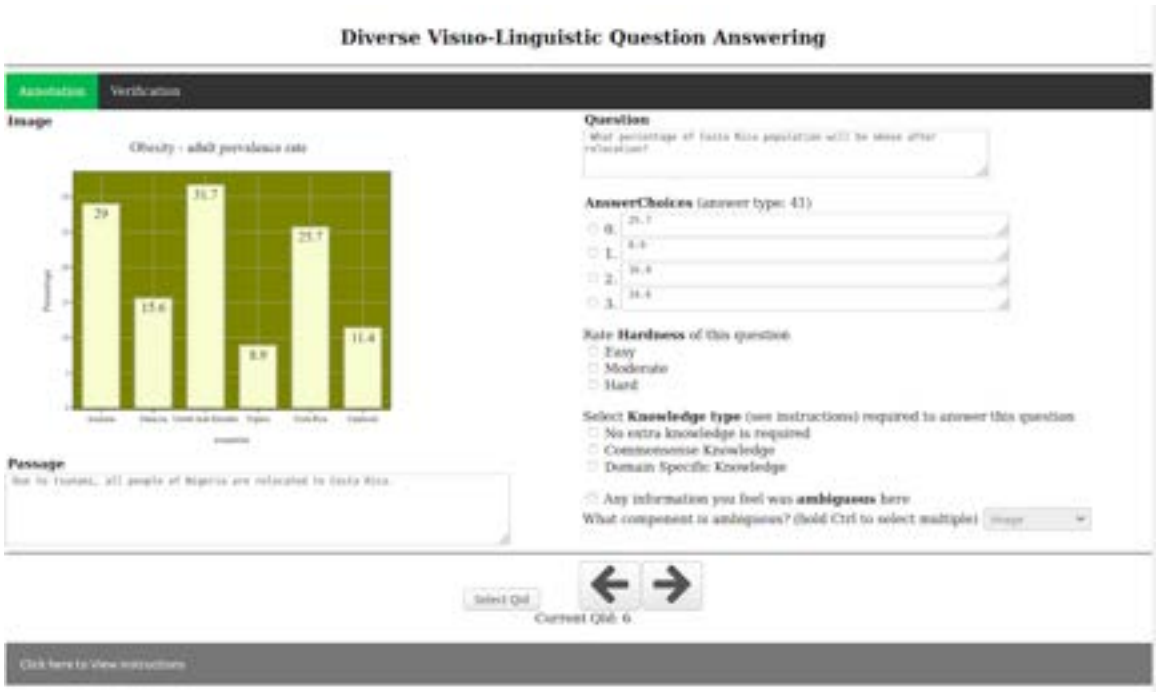


Figure 2.7: The Crowd Worker Interface Populated With an Example from VLQA During Initial Labeling. Specifically, <I, P, Q, A> are Shown on the Interface, and the User is Asked to Select the Correct Label (L), Assign the Appropriate Knowledge Category, Provide Their Judgment of the Hardness, and Report Ambiguity (if Any)

various biases in the dataset and obtain high accuracy without proper reasoning. To discourage such models, 3-level verification is employed over the full dataset to ensure the quality.

Firstly, for all collected image-passage pairs, human annotators quickly verify if a portion of the image and passage represent identical information. All such image-passage pairs are discarded from the dataset. Secondly, three baselines are prepared which ignore at least one modality (among image and passage), which are referred to as question-only, passage-only, and image-only models. This experiment is repeated three times, each baseline is posed with shuffled answer choices with a fixed seed and answers predicted by the system are logged. Samples that are answered correctly by

any unimodal baseline in all three trials are removed.

Finally, it is followed by another round of manual quality checks. The workers are instructed to attempt to answer a question based only on the image and then try to answer a question based only on the text passage. If a question can be answered using either modality, annotators are suggested to flag such questions. Finally, all bad samples are looked over and are either fixed or removed, on a case-by-case basis. Two rejected examples can be seen in Figure 2.8, with an explanation of the reason for removal.

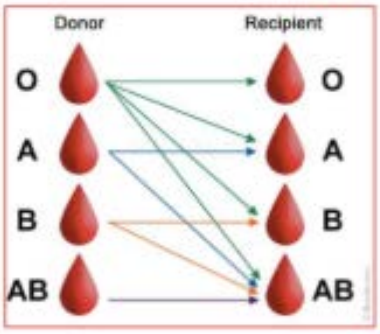
During the initial iteration, ∼12000 image-passage-QA pairs were collected. Further, in the annotation process, ∼700 samples were reported ambiguous, out of which ∼500 were removed and the remaining ∼200 were fixed and added back to the pool. By quality check process through unimodal baselines, ∼1900 samples were removed. In the verification stage, ∼350 samples were removed further. The resulting dataset consists of 9267 samples.

2.3.4 Dataset Statistics

In this section, analysis of the VLQA dataset is provided along several dimensions such as size and splits of the dataset, types of images, text-length, external knowledge required, reasoning abilities required, and manually annotated difficulty ratings. Table 2.1 and Table 2.2 provide a summary of relevant statistics for each dimension.

Multi-modal Contexts The final version of the VLQA dataset has 9267 unique image-passage-QA items. For each item, the multi-modal context is created by pairing images (roughly 10k collected) with the relevant text passages (roughly 9k retrieved or manually written).

(I) [0]



(P) The person who can receive blood from all groups is called a Universal Acceptor. The person who can donate blood to all other groups is called a Universal Donor. Joel has 'AB' blood group.

(Q) If Sarah is a Universal Donor, can she receive blood from Joel?

(A) a. Yes **b. No**

(Reason for rejection) If a person has memorized the concepts of Universal Donor and Universal Acceptor, the image is no longer necessary to answer the question

(I) [0]

(P) The figure shows healthcare spending as a percentage of GDP for all G7 countries.

(Q) Which of the following countries is not a G7 country?

(A) a. Canada **b. China** c. Japan d. USA

(Reason for rejection) Based on the simple web search, one can easily determine that which among the given choices is not a G7 country.

Figure 2.8: Example of Two Rejected VLQA Samples With an Explanation for the Rejection

Text-length Analysis There are three textual components in the VLQA dataset-passages, questions, and answers. An analysis of their length is conducted here. For each sample in the dataset, the count of tokens separated by whitespace along each component is computed and then averaged out across the dataset. The average passage length of 34.1 tokens indicates that in VLQA textual contexts are relatively smaller than Reading Comprehension tasks and in most cases, it contains precise context necessary for the joint reasoning. The average question length of 10.0 tokens is larger compared to most other VQA datasets (Hudson and Manning, 2019). Shorter answer lengths (1.7 tokens) suggest that most of the dataset questions have short answers, which provides inherent flexibility if someone wants to leverage generative models to solve this task. The dataset has a vocabulary size of 13259, contributed by all three kinds of textual components.

Image Types Images in VLQA are categorized into 3 major kinds: Natural Images, Template-based Figures, and Free-form Figures. Natural images incorporate day-to-day scenes, containing abundant objects and actions. Template-based figures are visuals that follow a common structure for information representation. Further, template-based figures are categorized into 20 sub-types like bar, pie, maps, tables, cycles, processes, etc. The images which neither fit in any templates nor are natural have been put into a free-form category (e.g., science experiments, hypothetical scenarios, etc.). In VLQA, it is also possible that the visual context has multiple related images to reason about.

Answer Types 4-way or 2-way image MCQ contains 4 and 2 images as plausible answer choices respectively, where the model needs to correctly pick the image best described by the passage and question. 4-way or 2-way text MCQ contains 4 and

2 alphanumeric text as plausible answer choices respectively, where the model needs to reason about a given image-text scenario and pick the most likely answer to the question. The 4-way Sequencing task assesses a model's capability to order 4 spatial or temporal events represented as a combination of images and text. Binary Classification (Yes/No or True/False) can be considered a fact-checking task where the truth value of a question is to be determined provided the image-passage context.

Knowledge and Reasoning Types 61% of VLQA items are observed to incorporate some commonsense or domain knowledge beyond the provided context. This missing knowledge has to be retrieved through the web. The remaining 39% samples can be answered through a simple join of information from a visuo-linguistic context. The following 10 reasoning types are most frequently needed to solve VLQA questions; conditional retrieval, math operations, deduction, temporal, spatial, causal, abductive, logical, and verbal reasoning. The VLQA samples also contain annotations based on whether it requires a single-step or multi-step inference to answer the question. The multi-step inference means that answering a question involves more than one reasoning types.

Difficulty Level Determining difficulty levels is a subjective notion therefore, an odd number of annotators are asked to rate VLQA items as 'easy', 'moderate', or 'hard' based on their personal opinions. A majority vote of all annotators determines the difficulty level of each question.

Dataset Splits VLQA contains 9267 items in <I, P, Q, A, L> format, with detailed classification based on figure types, answer types, reasoning skills, a requirement of external knowledge, and difficulty levels as explained above. The data is split in train-test-val (80-10-10%), ensuring the uniform distribution based on the above

Measure	Stats.
Multimodal Contexts	
Total #Images	10209
#Unique Text Passages	9156
#Questions	9267
Text-length Analysis	
Avg. Passage Length	34.1
Avg. Question Length	10.0
Avg. Answer Length	1.7
Vocabulary Size	13259
Image types	
Natural Images	4445
Templated Figures	3920
Free-form Figures	1854
Answer types	
4-way image MCQ	1172
4-way text MCQ	4647
4-way Sequencing	1088
2-way image MCQ	1088
Binary Classification (T/F or Yes/No)	1272

Table 2.1: VLQA Statistics and Diversity (MCQ Denotes Multiple Choice Question, Ext. Stands for External)

Knowledge/Reasoning types	
No Ext. Knowledge required	3145
Ext. Knowledge+Single-step Inference	2783
Ext. Knowledge+Multi-step Inference	2939
Difficulty Level (human annotated)	
Easy	4188
Moderate	2943
Hard	2136
Dataset Split	
Train (80%)	7413
Test (10%)	927
Validation (10%)	927

Table 2.2: (Continued) VLQA Statistics and Diversity

taxonomies. Note that the use of the metadata for model design is optional.

2.4 Experiments

This section includes a description of various baselines considered for the evaluation of the developed VLQA dataset. They are divided into four major kinds: heuristics (random and human performance), unimodal (question-only, passage-only, and image-only systems for data quality assurance), multimodal (pre-trained vision-language transformers), and a proposed multimodal system (fusion model based on off-the-shelf architectures).

2.4.1 Baseline Models

Human Performance A human evaluation is performed on 927 test samples with a balanced variety of questions by image types, answer types, knowledge/reasoning types, and hardness. Three in-house expert annotators take the test in isolation over multiple sessions. They select the answer choice for the given image+passage context and provide their judgment about the difficulty level of the question (easy/medium/hard). Their chosen answers are matched against the ground-truth answers, which turned out to be 84%.

Random Baseline VLQA dataset contains 4-way and 2-way multiple choice questions where each answer choice is likely to be picked with 25% and 50% chance respectively. Based on the answer-type distribution provided in Table 2.1, the chance of choosing an answer option randomly and that being correct is 31.36% across all question types.

Question-only, Passage-only, and Image-only Baselines Three unimodal baselines are used for automated quality assurance of VLQA (which are not trained or fine-tuned), to prevent models from exploiting bias in the data. Question-only, Passage-only, and Image-only models are implemented using RoBERTa (Liu *et al.*, 2019) fine-tuned on ARC (Clark *et al.*, 2018), ALBERT (Lan *et al.*, 2019) fine-tuned on RACE (Lai *et al.*, 2017) and LXMERT (Tan and Bansal, 2019) fine-tuned on VQA (Antol *et al.*, 2015a) respectively. The expectation here is that these unimodal baselines will demonstrate lower performance on the VLQA data, which will reassure the need for joint reasoning as per the task requirement. In other words, any system that ignores either the passage or image modality is likely to demonstrate poor performance on the VLQA dataset.

Vision-Language Transformers Recently, several attempts have been made to derive transformer-based pre-trainable generic representations for visual and text modalities. Among them, top-performing single-model architectures that support VQA tasks include VL-BERT (Su *et al.*, 2019), VisualBERT (Li *et al.*, 2019), ViLBERT (Lu *et al.*, 2019) and LXMERT (Tan and Bansal, 2019). Though all the aforementioned models have transformers (Vaswani *et al.*, 2017) as their backbone and support VQA tasks, they slightly differ in terms of the resources they have been pre-trained on, minor architectural variations, and support for other vision-language tasks. Their comparison is shown in Table 2.3, which is adapted from Lu *et al.* (2019).

For each model, two kinds of experiments are conducted- (1) use the pre-trained version of the model on the existing VQA dataset to obtain predictions on the VLQA dataset, and (2) take the pre-trained version of the model on the existing VQA dataset, further fine-tune it with VLQA data and then obtain predictions on VLQA samples. Most VQA systems only take one visual and one textual input. Hence, in the case of multiple images, they are composed into a single file. Similarly, the passage and questions are concatenated to form a single language input. Hyperparameters used for all four models are reported in Table 2.4.

2.4.2 Proposed Model: Modality Hopping-based Solver and Logical Entailment-based Reasoner

The aforementioned pre-trained V&L models have demonstrated strong performance over vision and language tasks separately. However, experimental results with respect to VLQA data (discussed in Section 2.5) indicate that they struggle with joint reasoning tasks. This suggests that the VLQA task is relatively harder compared to existing vision-language tasks due to the diversity of figures and additional textual components, demanding the need for better approaches to tackle multi-modal ques-

Model	Architecture and Pre-training	Supported Downstream Tasks
ViLBERT	One language transformer + one cross-modal transformer pre-trained on Conceptual captions	Visual question answering Visual commonsense reasoning Grounding referring expressions Image retrieval Zero-shot image retrieval
LXMERT	One language transformer + one vision transformer + one cross-modal transformer pre-trained on COCO caption + VG caption + VGQA + VQA + GQA	Visual question answering Natural language visual reasoning
VisualBERT	One cross-modal transformer pre-trained on COCO caption	Visual question answering Visual commonsense reasoning Natural language visual reasoning Grounding phrases
VL-BERT	One cross-modal transformer pre-trained on Conceptual captions + BooksCorpus + Wikipedia	Visual question answering Visual commonsense reasoning Grounding referring expressions

Table 2.3: Comparison of Vision-Language Transformers Used for the Experiments, Adapted From Lu *et al.* (2019)

Model Variations and Hyperparameters
VisualBERT
Ft_VQA: EP=20, BS=256, LR=1e-4, WD=1e-4
Ft_VLQA: BS=16, LR=2e-5, EP=15
VL-BERT
Ft_VQA: BS=32, LR=2e-5, EP=10
Ft_VLQA: BS=16, LR=1e-5, EP=10
ViLBERT
Ft_VQA: BS=32, LR=1e-5, EP=20, WR=0.1
Ft_VLQA: BS=32, LR=1e-5, EP=10
LXMERT
Ft_VQA: BS=32, LR=5e-5, EP=4
Ft_VLQA: BS=16, LR=5e-5, EP=8

Table 2.4: Hyperparameters Used for Experimentation: Ft_VQA Denotes Model Parameters Used While Training on Existing VQA Datasets. Ft_VLQA Denotes Parameters Used When Fine-tuning is Conducted Using VLQA
(Acronyms Used: BS- Batch Size, EP- Epochs, LR- Learning Rate, WD- Weight Decay, WR- Warmup Ratio)

tion answering. The VLQA challenge thus has the potential to open new research avenues spanning language and vision.

To address this, a modular and more interpretable method is proposed in this work. Specifically, given a sample from VLQA, based on the desired answer type ('4-way Image', '2-way Image', '4-way Sequencing', 'Binary Classification' or '4-

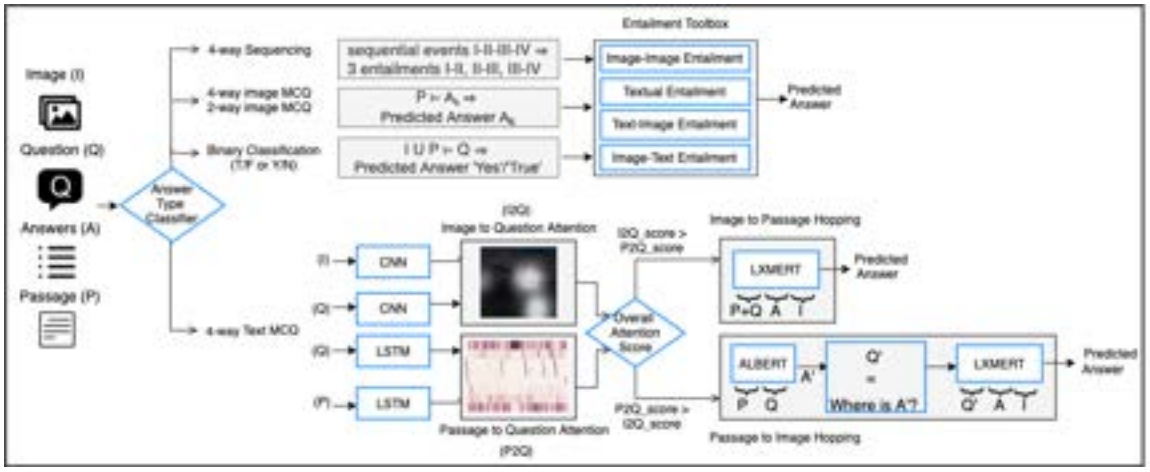


Figure 2.9: Proposed Method to Solve VLQA: Based on the Answer Type Classification, a Dataset Item is Either Solved Through a Sequence of Logical Entailment Operations or Modality Hopping

way Text'), there are two pathways- Modality Hopping (image-to-passage hop and passage-to-image hop) or Logical Entailment toolbox as shown in Figure 2.9. Specifically, 'answer type' annotations from the VLQA data are used to train a simple 5-class classifier. Note that the proposed model is not end-to-end.

Modality Hopping-based Solver

4-way text MCQ are solved using the modality hopping approach (lower half pipeline in Figure 2.9). First, image-to-question attention (I2Q) and passage-to-question attention (P2Q) scores are computed to determine which modality is important as a starting point for solving a question. I2Q is computed using Stacked Attention Network (SAN) (Yang *et al.*, 2016), which takes Convolution Neural Network (CNN) encoding of I and Q. Whereas, P2Q is computed using a variant of Bi-Directional Attention Flow (BiDAF) (Seo *et al.*, 2016) trained using Embeddings from Language Models (ELMo) (Peters *et al.*, 2018) over Long-Short Term Memory

(LSTM) encoding of Q and P.

A higher I2Q score suggests that Q has more overlap with I than P. Therefore, image modality should be used first and then incorporate passage to compute the answer. This is termed as an 'image-to-passage hop'. This is identical to a VQA model that takes an image and a question as input. Since P is an additional text component in this case, it is simply concatenated with the Q. This is implemented through pre-trained architecture LXMERT (Tan and Bansal, 2019) which is state-of-the-art on multiple choice VQA task.

Similarly, a higher P2Q score suggests that Q has more overlap with P than I. Therefore, passage modality should be used first and then incorporate an image to compute the answer. This is referred to as a 'passage-to-image hop'. This can be achieved by a machine comprehension model followed by a VQA model. ALBERT (Lan *et al.*, 2019) is used as a machine comprehension model that takes in P and Q to generate an open-ended response in the style of SQuAD (Rajpurkar *et al.*, 2016), which is referred to as A′. Following this, one would be required to locate A′ in image I. Therefore, a new question Q′ is composed as 'Where is A′?', where A′ is the answer generated by ALBERT (Lan *et al.*, 2019). Then LXMERT (Tan and Bansal, 2019) is employed, which takes image I, new question Q′, and original answer choices A to pick the most likely answer.

Logical Entailment-based Reasoner

For all other answer types, the Logical Entailment (upper half pipeline in Figure 2.9) pathway would be used to answer questions. An 'Entailment Toolbox' is created which consists of image-image, image-text (Xie *et al.*, 2019), text-image and textual entailment (Khot *et al.*, 2018) sub-modules and use them as required. The image-image and text-image entailment modules are developed by augmenting the

VisualCOPA dataset (Yeo *et al.*, 2018) and training a custom neural network over it.

4-way or 2-way image MCQ contains images as an answer choice, which is similar to an image selection task (Hu *et al.*, 2019). The goal here is to identify an image that best matches the description of P. Mathematically, determine $P \vdash {}^{3}A_k$ (i.e., text-image entailment) which yields maximum score, where k is 4 and 2 respectively for 4-way and 2-way image questions.

Binary classification can be considered as a fact-checking task where the truth value of a question is to be determined provided an image+passage context i.e., $P \cup I \vdash Q$. Textual entailment is used to determine $P \vdash Q$ and image-text entailment is used to determine $I \vdash Q$. If the score of both entailment modules is above 0.65, then it is determined as True, otherwise False.

4-way Sequencing task assesses a model's capability to order four spatial or temporal events. If four events I-II-III-IV (represented visually or textually) form a sequence, it is equivalent to three independent entailment inferences I-II, II-III, and III-IV. Among the answer choices, the sequence with maximum overall confidence is selected as an answer.

2.5 Results and Discussion

The VLQA dataset contains multiple-choice questions with exactly one correct answer, therefore exact match accuracy is used as an evaluation metric. The results for various uni-modal and multi-modal baselines (including the proposed method) along with the random and human performance on VLQA validation and test sets are summarized in Table 2.5. It is apparent that pre-trained vision-language models

[3] $\vdash$ is the symbolic representation of entailment

Method	Test(%)	Val(%)
Human	84.00	–
Random	31.36	31.36
Question-only: RoBERTa$_{ARC}$	28.56	29.42
Passage-only: ALBERT$_{RACE}$	30.16	30.25
Image-only: LXMERT$_{VQA}$	29.48	30.56
Vision-Language		
VL-BERT	35.92	34.60
VisualBERT	33.17	34.17
ViLBERT	34.70	35.25
LXMERT	**36.41**	37.82
Proposed Model	**39.63**	40.08

Table 2.5: Performance Benchmarks Over Test Set of the VLQA Task and Corresponding Validation Results

fail to solve a significant portion of the VLQA items and they are barely better than the random baseline. The proposed method slightly outperforms other vision-language models and is more interpretable from an analysis perspective as it makes the reasoning process modular. As discussed previously, the lower performance of question-only, image-only, and passage-only baselines is a way to reassure the joint reasoning requirement in VLQA.

On the other hand, the reported accuracy for human evaluation on the VLQA test set is 84.0%. This result indicates that there is significant room for improvement in the multi-modal reasoning capabilities of existing vision-language models in comparison

Underlying reason for incorrect answer provided by test-taker	**#incorrect/148 (%incorrect)**
Lacked necessary knowledge	27 (18.2%)
Misunderstood the provided info	47 (31.7%)
Mistake in deduction/calculation	63 (42.5%)
Felt that data item is ambiguous	11 (7.4%)

Table 2.6: Classification of Incorrectly Predicted Answers in Human-evaluation of VLQA Test Data

with humans. For 148 questions that were wrongly predicted by humans, a manual study was conducted to investigate the reasons behind the failures. This included errors made by the test takers, incorrect interpretation or ambiguity in the provided information, and the lack of necessary knowledge required to answer questions. The distribution based on the aforementioned failure cases is summarized in Table 2.6. In a majority of cases, it was a mistake on the human annotator's end during the deduction process or mathematical computation.

In conclusion, multi-modality brings both pros and cons while developing new AI benchmarks. The presence of multiple modalities provides natural flexibility for varied inference tasks, simultaneously making the reasoning process more complex as information is now spanned across them and requires cross-inferencing. In this work, the importance of joint reasoning capability over image-text context is emphasized, inspired by how humans use such an ability in day-to-day tasks. To support the development of AI systems that can perform multi-modal reasoning, a Visuo-Linguistic Question Answering (VLQA) Dataset (Sampat *et al.*, 2020b) is proposed. The proposed dataset has important distinctions from other existing VQA datasets. Firstly, it

incorporates a text passage that contains additional contextual information and does not simply reflect the information already conveyed in images. Secondly, it offers various figure types including natural images, template-based images (such as charts), and free-form images. Thirdly, it tests diverse reasoning capabilities, including cross-inferencing between visual and textual modalities. The experimentation with respect to several recent models including pre-trained vision-language transformers indicates that they highly rely on image-text similarity to learn representations and struggle to perform joint inference over heterogeneous modalities.

Chapter 3

HYPOTHETICAL REASONING ABOUT ACTIONS: A SPECIAL CASE OF VLQA

3.1 Motivation

In 2014, Michael Jordan (Distinguished Professor at the University of California, Berkeley) in an interview (Gomes, 2014) said that "*Deep learning is good at certain problems like image classification and identifying objects in the scene, but it struggles to talk about how those objects relate to each other, or how a person/robot would interact with those objects. For example, humans can deal with inferences about the scene: what if I sit down on that?, what if I put something on top of something? etc. There exists a range of problems that are far beyond the capability of today's machines.*"

While this interview was six years ago, and since then there has been a lot of progress in deep learning and its applications to visual understanding. Additionally, a large body of visual question answering (VQA) datasets have been compiled (Antol *et al.*, 2015b; Ren *et al.*, 2015a; Hudson and Manning, 2019) and many models have been developed over them. Still, the aforementioned "inferences about the scene" issue stated by Jordan remains largely unaddressed.

In most existing VQA datasets, scene understanding is holistic and questions are centered around information explicitly present in the image (i.e. objects, attributes, and actions). As a result, advanced object detection and scene graph techniques have been quite successful in achieving good performance over these datasets. However, provided an image, humans can speculate a wide range of implicit information. For

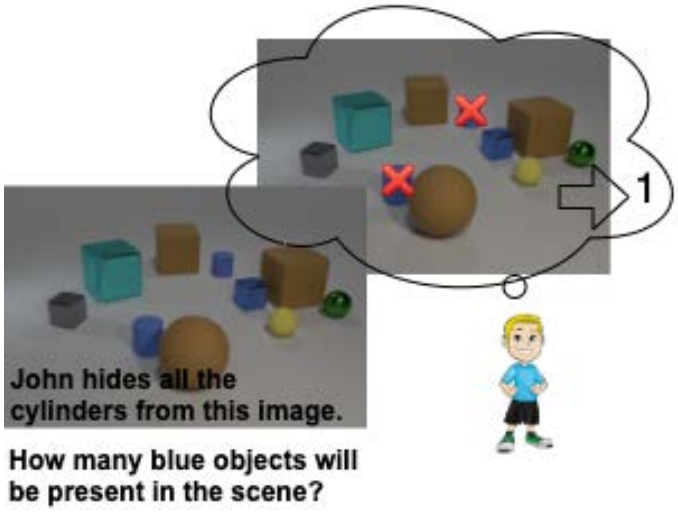


Figure 3.1: Motivation for the Proposed Hypothetical Reasoning Task in the Vision-Language Domain: an Example Demonstrating How Humans can Perform Mental Simulations and Reason Over the Resulting Scenario

example, the purpose of various objects in a scene, speculation about events that might have happened before, visualize numerous imaginary situations that might be possible and predict possible future outcomes, intentions of a subject to perform particular actions, and many more.

Among the above, the ability to imagine taking specific actions and simulating probable results without actually acting or experiencing is an important aspect of human cognition. Consider the example in Figure 3.1, where a person is shown a visual scene and told that- John hides all the cylinders from that image. Provided this, the person would imagine performing this action and reason that two cylinder objects will no longer be in the scene. Now, if someone asks a question 'How many blue objects are present in the scene?', the person would answer '1' considering the updated object configuration. In other words, a small blue cube is the only blue object left in the image after John hides all the cylinders.

Thus, having autonomous systems equipped with similar capabilities will be beneficial and advance the state-of-the-art of AI research. This is particularly useful for robots performing on-demand tasks in safety-critical situations or navigating through dynamic environments, where they imagine possible outcomes for various situations without executing instructions directly. Motivated by the above, a challenging task is proposed to bridge the gap between state-of-the-art AI and human-level cognition. The main contributions of this work are as follows;

- A novel task is proposed where given a state of the world (in a visual form) and an action (described in a textual form) to be performed over the world, an AI model has to reason about the effects of actions on the world.
- A large-scale synthetic dataset is developed for this task, which is referred to as CLEVR_HYP i.e. VQA task with hypothetical actions performed over images in CLEVR (Johnson *et al.*, 2017a) style.
- Direct extensions of recent VQA and NLQA (Natural language QA) solvers are evaluated on this dataset and a couple of new baseline models are proposed.
- Through analysis and ablations, insights on the capability of diverse architectures to perform joint reasoning over image-text modality are provided.

Note that, the focus here is to develop neural models that can reason about the effects of actions given a visual-linguistic context. There is a dedicated research direction that tackles similar tasks by learning intuitive physics principles, which is beyond the scope of this work.

3.2 CLEVR_HYP Task Overview

Figure 3.2 gives a glimpse of the CLEVR_HYP task. The system is provided three inputs- Image(I), Action Text (T_A), and Hypothetical Question (Q_H). The system is expected to select Answer (a) to the question, which requires reasoning about the effects produced by performing action over the objects in the image. Each component is described in detail below;

3 Inputs in the CLEVR_HYP task:

1. Image(I): It is a given visual that contains 3-10 objects. Each object has four attributes- color, shape, size, and material, using which objects can be identified/referred.

2. Action Text (T_A): It is a natural language text describing various actions performed over the current scene. The action can be one of the following four kinds:

(i) **Add** new object(s) to the scene

(ii) **Remove** object(s) from the scene

(iii) **Change** attribute(s) of the object(s)

(iv) **Move** object(s) within the scene, either in plane i.e. left/right/front/back, or out of plane i.e. move one object on top of another object.

For simplicity, it is assumed that any object can be put on another object regardless of its size, material, or shape.

3. Question about Hypothetical Reasoning (Q_H): It is a natural language query that tests various visual reasoning abilities of the system after simulating the effects of actions described in T_A. There are five possible reasoning types similar to CLEVR (Johnson *et al.*, 2017a);

I: 

Example 1:

$\mathbf{T}_A$: Paint the small green ball with cyan color.

$\mathbf{Q}_H$: Are there equal number of yellow cubes on the left of the purple object and cyan spheres?

a: Yes

Example 2:

$\mathbf{T}_A$: Add a brown rubber cube behind the blue sphere that inherits its size from the green object.

$\mathbf{Q}_H$: How many things are either brown or small?

a: 6

Example 3:

$\mathbf{T}_A$: John moves the small red cylinder on the large cube that is to the right of the purple cylinder.

$\mathbf{Q}_H$: What color is the object that is at the bottom of the small red cylinder?

a: Yellow

Figure 3.2: Three Examples From CLEVR_HYP Dataset: Given Image (I), Action Text (T_A), and Question (Q_H), the Model Predicts an Answer (a). The Task is to Understand Possible Perturbations in I With Respect to Various Action(s) Performed as Described in T_A. Questions Test Various Reasoning Capabilities of a Model With Respect to the Changes Caused by the Action(s).

(i) **Counting** objects fulfilling the given criteria

(ii) **Verify existence** of certain objects

(iii) **Query attribute** of a particular object

(iv) **Compare attributes** of two objects

(v) **Integer comparison** of two object sets (same, larger, or smaller)

Output of the CLEVR_HYP task:

Answer (a): It is an answer to the Question (Q_H), which can be considered as a 27-way classification over object attributes (8 colors + 3 shapes + 2 sizes + 2 materials), numerals (0-9) and booleans (yes/no).

The CLEVR_HYP task can be considered as a special case of Visuo-Linguistic Question Answering (introduced in Chapter 2) as per the formulation below:

$$\begin{aligned} &\text{VLQA Task Formulation: } f_\theta(I, P, Q) \rightarrow \mathtt{a} \in A \\ &\text{CLEVR\_HYP Task Formulation: } f_\theta(I, T_A, Q_H) \rightarrow \mathtt{a} \in A \end{aligned} \tag{3.1}$$

Here, I is a rendered visual with a pre-defined set of object attributes (as shown in the example in Figure 3.2); T_A is a special case of P which is limited to the description of actions that are to be performed over the objects in the image; Q_H is a special case of Q which is limited to visual reasoning about objects, their properties, and elementary math (counting and comparison in particular).

3.3 CLEVR_HYP Dataset Creation

As discussed in Section 1.3, there are no existing benchmarks useful for the evaluation of hypothetical reasoning abilities in AI systems. As observed in Chapter 2, the human-in-the-loop style dataset creation process is expensive, time-consuming,

and not scalable in nature. Moreover, failure in object recognition is a major cause of error propagation. Therefore, in this work, a synthetic dataset creation process is opted which addresses the above limitations to some extent.

3.3.1 Data Collection

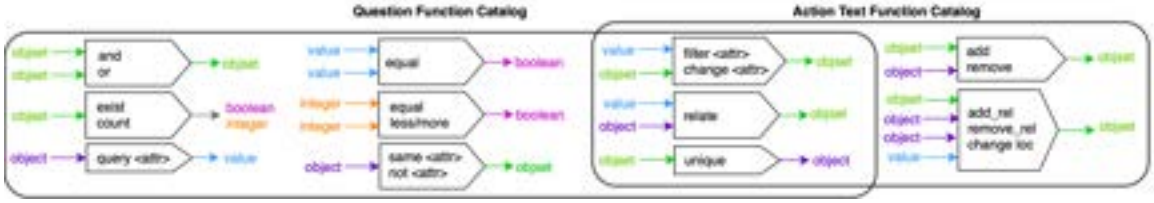
The CLEVR_HYP dataset is automatically generated using the procedure shown in Figure 3.3b. The synthetic dataset creation process allows data generation at scale, controls biases, and eliminates the need for expensive human annotations. Following is a brief description of how each of the inputs and ground-truth answers are obtained;

1. Image (I): Each image in the dataset contains 3-10 randomly selected 3D objects rendered using Blender (Blender Online Community, 2019) following CLEVR (Johnson *et al.*, 2017a). Each object has four attributes- color, shape, size, and material. Each attribute takes a pre-defined set of values listed as per Table 3.1. Additionally, the objects can be referenced using five relative spatial relations (left, right, in front, behind, and on). The scene graphs containing all ground-truth information about scenes are provided during training time, which can be considered as a visual oracle for a given image.

Attribute	**Possible values in CLEVR_HYP**
Color	gray, blue, brown, yellow, red, green, purple, cyan
Shape	cylinder, sphere or cube
Size	small or large
Material	metal (shining object) or rubber (matte object)

Table 3.1: Object Attributes in CLEVR_HYP Scenes

(a) Action and question function catalog for CLEVR_HYP, extended from CLEVR (Johnson *et al.*, 2017a)



(b) Dataset creation pipeline

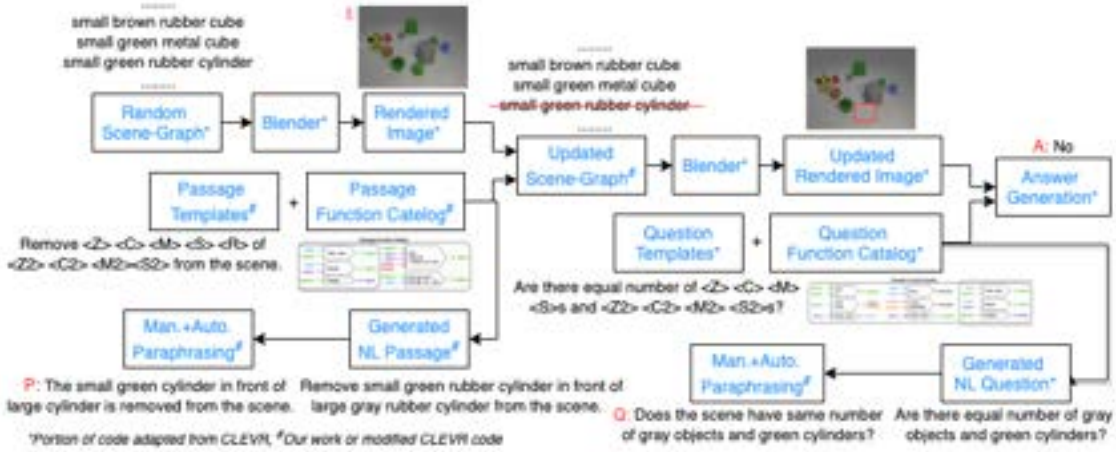


Figure 3.3: CLEVR_HYP Dataset Creation Process With an Example and Function Catalogs Used for Ground-truth Answer Generation

2. Action Text ($\mathbf{T}_A$): To generate action text (add, remove, change, or move), manually written templates are leveraged. For example, to represent action involving a change in the attribute of object(s) to a given value, a template 'Change the <A> of <Z><C><M><S> to <V>' is used. Where <A>, <Z>, <C>, <M>, <S>, <V> are placeholders for the attribute, size, color, material, shape and a value of attribute respectively. Each action text in the CLEVR_HYP is associated with a functional program which if executed on an image's scene graph, yields the new scene graph that simulates the effects of actions.

Functional programs for action texts are built from the basic functions that cor-

respond to elementary action operations (right part of Figure 3.3a). For the above change attribute action template, the equivalent functional program can be written as; 'change_attr(<A>, filter_size(<Z>, filter_color(<C>, filter_material(<M>, filter_shape(<S>, scene()))))), <V>)'. It essentially means, first, filtering out the objects with desired attributes from the scene and then updating the value of their current attribute A to value V.

3. Question about Hypothetical Reasoning (Q_H): Similar to action texts, there are templates and corresponding programs for questions. Functional programs for questions are executed on the updated scene graph of the image (after incorporating the effects of the action text) and yield ground-truth answers for the given questions. Functional programs for questions are made of primitive functions shown in the left part of Figure 3.3a).

Here, an action text and a question are kept as separate inputs for the purpose of simplicity and keeping the focus on building solvers that can do mental simulations. It is possible to create a simple template like '<Q_H> if <proper-noun/pronoun> <T_A>?' or 'If <proper-noun/pronoun> <T_A>, <Q_H>?' if someone wishes to process action and question as single text input. For example, 'How many things are the same size as the cyan cylinder if I add a large brown rubber cube behind the blue object.' or 'If I add a large brown rubber cube behind the blue object, how many things are the same size as the cyan cylinder?'. However, having them together adds further complexity on the solver side as it first has to figure out what actions are performed and what is the question.

As discussed above, ground-truth object information (as a visual oracle) and machine-readable form of questions as well as action texts (oracle for linguistic components) are made available during the training phase. This information can be used

to develop models that can process semi-structured representations of image/text or for explainability purposes (to precisely know which component of the model is failing).

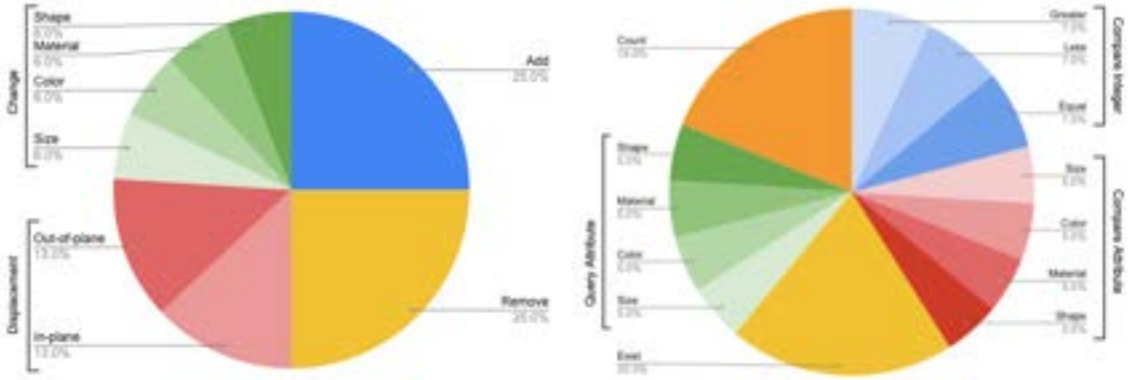
3.3.2 Ensuring Dataset Integrity

As described above, the images are rendered from scene graphs, and templates are used to generate text components. In this way, it is possible to keep track of each component and have scripts to automatically generate the ground-truth answers. This process is completely automated and no manual annotations are required. Following is a brief description of measures taken to control the bias and use of paraphrasing for linguistic diversity of the dataset;

Bias Control: For each image, five action texts are generated (one for each add, remove, move in-plane, move out-of-plane and change attribute action). For each action type, five questions are produced (one for each count, exist, compare integer, query attribute, and compare attribute). Hence, 25 (i.e. 5*5) unique action text-question pairs are obtained for each image, covering all actions and reasoning types in a balanced manner as shown in Figure 3.4a (referred to as the Original partition in subsequent sections). However, it leads to a skewed distribution of answers as observed from Figure 3.4b. Therefore, a version of the dataset is created (referred to as the Balanced partition in subsequent sections) consisting of 67.5k samples where all answer choices are equally likely as well.

Paraphrasing: In order to create a challenging dataset from the linguistic point of view and to prevent models from overfitting on templated representations, three measures are taken- noun synonyms, object name paraphrasing, and sentence-level

(a) Distribution based on action types and question types



(b) Distribution of answer types

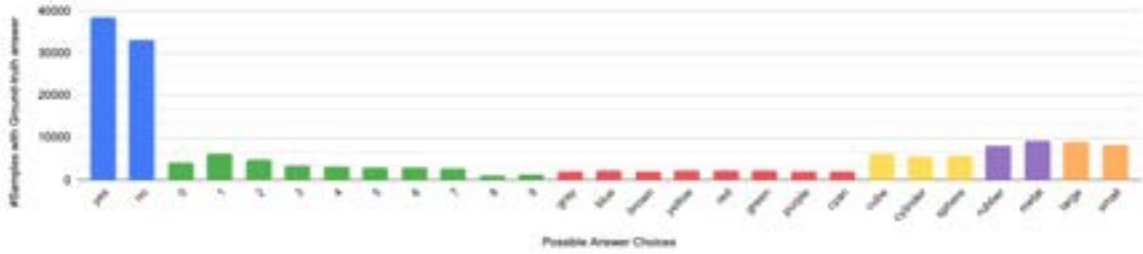


Figure 3.4: Visualization Showing Distribution by Actions, Questions, and Answers in Original_Train Partition of the CLEVR_HYP

paraphrasing. For noun synonyms, a pre-defined dictionary (such as cube∼block, sphere∼ball, and so on) is used to randomly replace nouns in the generated action texts and questions. Further, all attribute compositions are generated that refer to a given object uniquely in the scene and one among them is randomly used, which is called object name paraphrasing. Finally, for sentence-level paraphrasing, the Text-To-Text Transfer Transformer (T5) model (Raffel *et al.*, 2020) fine-tuned over positive samples from Quora Question Pairs (QQP) dataset (Iyer *et al.*, 2017) is used to paraphrase questions. For the action text paraphrasing, the Fairseq (Ott *et al.*, 2019) model variation with round-trip translation and a mixture of experts (Shen *et al.*, 2019) is used.

3.3.3 Dataset Statistics

Using the process mentioned in Figure 3.3b, 175k image-action text-question samples were collected (referred to as the Original partition) in the first iteration. However, this data had a skewed distribution of answers. In order to make the dataset less prone to biases, a subset is created where the distribution of actions, reasoning types, and answers is uniform. This is referred to as a Balanced partition, which consists of 67.5k samples.

Additionally, two challenge test sets- 2HopActionText (2HopT$_A$) and 2HopQuestion (2HopQ$_H$) are created to test the generalization capability of the trained models. Both these test sets have 1500 image-action text-question samples. In the 2HopT$_A$ test, the model is required to perform two different actions one after the other on the scene. For example, 'Add a small blue metal cylinder to the right of the large yellow cube and remove the large cylinder from the scene.' and 'Move the purple object on top of the small red cube then change its color to cyan.'. In the 2HopQ$_H$ test, the model requires understanding questions involving logical operations 'and', 'or' and 'not'. For example, 'How many objects are either red or cylinder?' and 'Are there any rubber cubes that are not green?'.

Table 3.2 summarizes various partitions of the dataset, including their size and diversity. For images, an average number of objects is computed. For the balanced partition, the number of images is lower in comparison with the original partition, but a higher number of objects per image on average. This is likely due to the need to accommodate integers 4-9 as ground-truth answers.

For textual components, average lengths (number of tokens separated by whitespace) and count of unique utterances are considered as a measure of diversity. The original partition of the resulting dataset has 80% and 83% unique action text and

Split	**#I**	**Avg. #Obj**	**#T$_A$**	**Uniq. #T$_A$**	**Avg. T$_A$ Len.**	**#Q$_H$**	**Uniq. #Q$_H$**	**Avg. Q$_H$ Len.**
Orig_Train	5k	6.4	25k	20.7k	12.8	125k	103.7k	22.6
Orig_Val	1k	6.7	5k	3.8k	12.8	25k	20.9k	23.1
Orig_Test	1k	6.4	5k	3.6k	12.6	25k	20.7k	22.8
Bal_Train	5k	7.6	25k	21.1k	12.8	67.5k	58.2k	20.3
Bal_Val	1k	7.6	5k	3.9k	12.7	13.5k	11.5k	20.7
Bal_Test	1k	7.5	5k	3.7k	12.6	13.5k	11.4k	20.4
2HopT_A	500	6.4	300	260	18.6	1.5k	1.25k	22.8
2HopQ_H	500	6.4	300	220	12.6	1.5k	1.37k	29.3

Table 3.2: CLEVR_HYP Dataset Splits and Statistics
(#- Number of, k- Thousand, Orig.- Original, Bal.- Balanced, and Uniq.- Unique)

questions respectively. For the balanced partition, length and unique utterances for action text are nearly the same as the original partition but for questions, it decreases. Questions in the original partition have been observed to enforce more strict and specific object references (such as small red metal cubes) compared to a balanced partition (small cubes, red metal objects, etc.), reducing the average length and uniqueness. It is intuitive for 2Hop partitions to have higher average length and uniqueness for T_A and Q_H respectively. This shows that despite having created this dataset from templates and rendered images with a limited set of attributes, it is still fairly challenging.

3.4 Experiments

3.4.1 Baseline Models

Models that attempt to tackle the CLEVR_HYP task have to address four key challenges;

(i) Understand hypothetical actions and questions in complex natural language,

(ii) Correctly disambiguate the objects of interest and obtain the structured representation of various modalities (i.e. scene graphs or functional programs if required by the solver),

(iii) Understand the dynamics of the world based on the actions performed over it,

(iv) Perform various kinds of reasoning to answer questions.

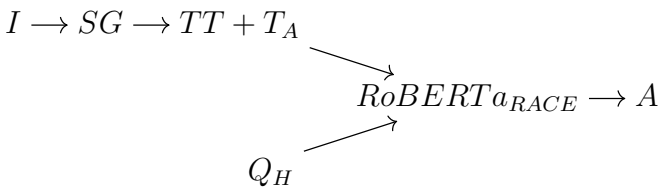
Therefore, this task is more challenging in comparison with standard VQA where models look up information that is always present in the image (i.e. objects, attributes, and actions). A detailed description of four baseline solvers employed for experimentation in this work is provided below. Figure 3.5 provides a graphical visualization of each baseline.

Baseline 1- Machine Comprehension using RoBERTa: To evaluate the hypothetical VQA task through the text-only model, images are converted into the templated text (equivalent to a deep caption of the image) using the scene graphs. The templated text contains two kinds of sentences; one describing properties of the objects i.e. 'There is a <Z> <C> <M> <S>', the other one describing their relative spatial location i.e. 'The <Z> <C> <M> <S> is <R> the <Z1> <C1> <M1> <S1>'. For example, 'There is a small green metal cube.' and 'The large yellow

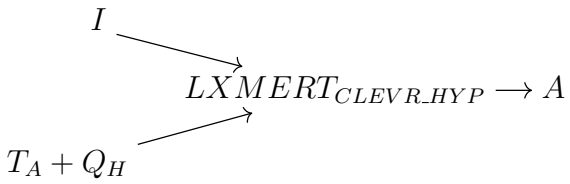
Symbols used:

I: Image	SG: Scene Graph	TT: Templated Text
T_A: Action Text	Q_H: Hypothetical Question	A: Answer
FP: Functional Program	$'$: Updated Modality	

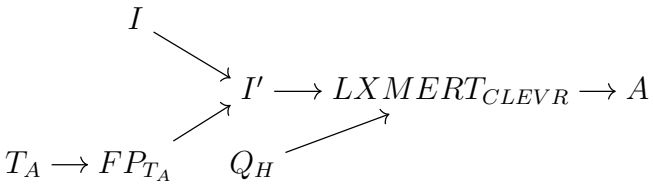
Baseline 1:



Baseline 2:



Baseline 3:



Baseline 4:

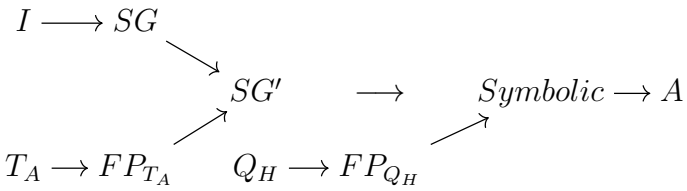


Figure 3.5: Graphical Visualization of Baseline Models Over CLEVR_HYP Which are Implemented in This Work

rubber sphere is to the left of the small green metal cube'. Then, the aforementioned templated description of the image is concatenated with the action text to create a reading comprehension passage and provided along with the question to the model. For this purpose, state-of-the-art machine comprehension baseline RoBERTa_large (Liu *et al.*, 2019) fine-tuned on the RACE dataset (Lai *et al.*, 2017) is used. The hyperparameters set during the fine-tuning process on the RACE dataset includes epochs=5, learning rate=1e−05, batch size=2, update frequency=2, dropout=0.1, optimizer=adam with epsilon=1e−06.

Baseline 2- Visual Question Answering using LXMERT: Proposed by Tan and Bansal (2019), LXMERT is one of the best transformer-based pre-trainable visual-linguistic representations that support VQA as a downstream task. Typical VQA systems take an image and a language input. Therefore, to evaluate CLEVR_HYP in VQA style, action text is concatenated with the question to form a single text input. Since LXMERT is pre-trained on the natural images, it is fine-tuned on CLEVR_HYP dataset with hyperparameters epochs=4, learning rate=5e−05, batch size=8 and then used to predict the answer.

Baseline 3- Text-editing Image (TIE) Baseline: In this method, the hypothetical reasoning process is divided into two sub-tasks; first, learning to generate an updated image (such that it has incorporated the effects of actions) and second, performing visual question answering with respect to the updated image. The first sub-task is implemented using Text Image Residual Gating proposed by Vo *et al.* (2019). However, there are two important distinctions from their work; Their focus is on retrieval from the given database. Their objective is modified here and a text-adaptive encoder-decoder with residual connections is developed to generate the new

image. Also, editing instructions in their CSS dataset were quite simple. For example, 'add red cube' and 'remove yellow sphere'. With such open-ended instructions, one can add the red cube anywhere in the scene. In this work, one is required to precisely place objects according to their relative spatial references (on left/right/front/behind). Once the updated image is obtained, it is provided to LXMERT (Tan and Bansal, 2019) fine-tuned over the CLEVR (Johnson *et al.*, 2017a) along with the question to predict the answer.

Baseline 4- Scene Graph Update (SGU) Model: This method also divides the hypothetical reasoning process into two sub-tasks. Contrary to directly manipulating images in the first sub-task like Baseline 3, this method employs effect prediction as a graph-editing operation conditioned on the action text. This is an emerging research direction to deal with changes in the visual modality over time or with new sources of information, as observed from recent parallel works such as Chen *et al.* (2020a); He *et al.* (2020).

Towards this end, the segmentation mask of the image is obtained in order to detect using the Mask R-CNN (He *et al.*, 2017). On top of the segmentation mask, there is a classifier to predict object attributes (color, material, size, and shape) with an acceptance threshold of 0.9. The segmentation mask of each object along with the original image is then passed through ResNet-34 (He *et al.*, 2016) to extract precise 3D coordinates of the object. Predictions from all the above modules are then combined to obtain the structured scene graph for the input image. Then seq2seq model with attention (Johnson *et al.*, 2017b) is used to obtain functional programs for action text and question. A reinforcement learning algorithm is leveraged to update scene graphs provided by the functional program for the action text. The parameters used for this algorithm include learning rate=1e−05, 1M iterations with early stopping,

and batch size=32. The reward function is associated with the ground-truth program executor and generates a reward if the prediction exactly matches with ground-truth execution. Once the updated scene representation is obtained, the neural-symbolic model proposed by Yi *et al.* (2018) is used to answer the question. The supervised pre-training of Yi *et al.* (2018)'s model was done with the learning rate=7e−04, number of iterations=20k, batch size=32 and further fine-tuning with the learning rate=1e−05, at most 2M iterations with early stopping, and batch size=32. It is notable that Yi *et al.* (2018) achieved near-perfect performance on the CLEVR QA task in addition to being fully explainable.

3.5 Results and Discussion

In this section, the performance of four baseline methods developed over the CLEVR_HYP task (described above) is discussed. The dataset is formulated as a classification task with exactly one correct answer, therefore a standard accuracy metric is used for evaluation. A comparison with random chance baseline and human performance on this task is also discussed. Further, for each baseline, an analysis of its performance according to question types and action types is conducted.

Random The QA task in the CLEVR_HYP dataset can be considered as a 27-class classification problem. Each answer choice is likely to be picked in the balanced partition of the dataset with a probability of 1/27. Therefore, the performance of the random baseline is 3.7%.

Human Performance A human evaluation is conducted with respect to 500 samples from the CLEVR_HYP dataset. Accuracy of human evaluation on original, 2HopA_T and 2HopQ_H test sets turned out to be 98.4%, 96.2%, and 96.6% respec-

Overall Performance of baseline models over various test sets of CLEVR_HYP								
Original Test				*Balanced Test*			
BL1	BL2	BL3	BL4	BL1	BL2	BL3	BL4
57.2	63.9	64.7	**70.5**	55.3	65.2	**69.5**	68.6
2HopTA Test				*2HopQH Test*			
BL1	BL2	BL3	BL4	BL1	BL2	BL3	BL4
53.3	49.2	55.6	**64.4**	55.2	52.9	58.7	**66.5**

Table 3.3: Overall Performance of Baseline Models Developed for CLEVR_HYP Task, Corresponding to Four Test Sets in the Dataset (BL1 - Machine Comprehension using RoBERTa, BL2 - VQA using LXMERT, BL3 - Text-editing Image Model, BL4 - Scene Graph Update Model)

tively.

Baseline Performance Quantitative results for the baseline models discussed in Section 3.4 are summarized in Table 3.3. Among the methods described above, the scene graph update model (baseline 4) has the best overall performance on original test data with 70.5% accuracy. The text-editing model (baseline 3) is top-performing over the balanced test set, however, it demonstrates poor generalization capability (in 2HopA_T and 2HopQ_H tests). The CLEVR_HYP task requires models to reason about the effect of hypothetical actions taken over images. As the LXMERT (baseline 2) is not directly trained for this objective, it struggles to do well on this task. The reason behind the poor performance of the text-only baseline (baseline 1) is due to its limitation in incorporating detailed spatial locations into the templates that are

Performance break-down by Action Types					
	Original Test			*2HopA_T Test*	
	BL3	BL4		BL3	BL4
Add	58.2	65.9	Add+Remove	53.6	63.2
Remove	89.4	88.6	Add+Change	55.4	64.7
Change	88.7	91.2	Add+Move	49.7	57.5
Move(in-plane)	61.5	69.4	Remove+Change	82.1	85.5
Move(on)	53.3	66.1	Remove+Move	52.6	66.4
			Change+Move	53.8	63.3

Table 3.4: Performance Breakdown by Action Types for BL3 (Text-editing Image Baseline) and BL4 (Scene Graph Update Baseline)

used to convert an image into a machine comprehension passage.

Analysis by Action Types and Question Reasoning Types Two of the proposed models (baseline 3 and 4) are transparent to visualize intermediate changes in the scene after performing actions. Their ability to understand actions and make appropriate changes in the scene are further analyzed as per Table 3.4. For the scene graph method, the ground-truth functional program is compared with the model-generated functional program for each sample, and their exact match accuracy is computed. For the text-editing image method, the ground truth scene graph after action execution and the scene graph generated using the model-edited image are compared. For attributes, exact matching is sought whereas for object positions, their relative spatial locations need to match.

The results from Table 3.4 indicate that both the scene graph and text-editing

Performance break-down by Reasoning Types					
	Original Test			*2HopQ$_H$ Test*	
	BL3	BL4		BL3	BL4
Count	60.2	74.3	And	59.2	67.1
Exist	69.6	72.6	Or	58.8	67.4
Compare Integer	56.7	67.3	Not	58.1	65.0
Compare Attribute	68.7	70.5			
Query Attribute	65.4	68.1			

Table 3.5: Performance Breakdown by Question Reasoning Types for BL3 (Text-editing Image Baseline) and BL4 (Scene Graph Update Baseline)

models do quite well on 'remove' and 'change' actions whereas struggle relatively when new objects are added or existing objects are moved around. The observation is consistent when multiple actions are combined. In other words, actions remove+change can be performed with maximum accuracy whereas other combinations of actions accomplish relatively lower performance. It leads to the conclusion that understanding the effect of different actions is of varied complexity.

A similar analysis is conducted with respect to question reasoning types. After an updated image or a scene graph is obtained as a result of performing an action, how accurately the models can predict correct answers for various question types. The performance breakdown of baseline 3 and 4 with respect to reasoning types is shown in Table 3.5. Both the models demonstrate relatively higher performance over counting, existence, and attribute query question types than the comparison questions. The scene graph update and text-editing methods show a performance drop of 6.1% and 9.1% respectively when multiple actions are performed on the scene. However, there is

less performance gap for models on 2HopQ$_H$ compared to the 2HopA$_T$ test set. This suggests that both models can better generalize while performing complex logical reasoning than understanding the effects of complex actions.

In summary, this work presents CLEVR_HYP (Sampat *et al.*, 2021), a dataset to evaluate the ability of VQA systems where hypothetical actions are to be performed over the given image. This dataset is created by extending the data generation framework of CLEVR (Johnson *et al.*, 2017a) that uses synthetically rendered images and templates for reasoning questions. This dataset is challenging because rather than asking questions about objects already present in the image, it poses questions to models about what would happen in an alternative world where changes have occurred. The dataset provides a visual oracle for images and textual oracles for hypothetical actions as well as questions to facilitate the development of models that systematically learn the underlying action reasoning process. Several baselines are developed based on existing vision and language models to tackle the hypothetical reasoning task. The results demonstrate that the models are able to perform reasonably well (70.5%) on the limited number of actions and reasoning types, but struggle with more complex scenarios. While neural models have achieved almost perfect performance on CLEVR and considering human performance as an upper bound (98%), there is a lot of room for improvement on the CLEVR_HYP dataset. Improvements over this dataset and its extensions by relaxing various constraints (including a larger variety of actions, attributes, and reasoning types) will lead a path toward better vision+language models that can reason beyond pixels.

Chapter 4

ACTION REPRESENTATION LEARNER (ARL): AN IMPROVED MODEL FOR HYPOTHETICAL REASONING ABOUT ACTIONS

4.1 Motivation

Humans interact with their environment to accomplish desired goals. Object manipulation (i.e., performing 'actions' over the objects) is a fundamental concept that makes this interaction possible. In other words, actions in their simplest form have the power to change the state of a world and hence play a vital role in enabling humans to perform day-to-day tasks. As autonomous agents are being developed that would assist humans in performing everyday tasks, such agents would be required to interact with complex environments. Hence, the development of autonomous agents that can perform actions to effectively manipulate objects and understand corresponding effects is of great importance. As a result, Reasoning about Action and Change (RAC) has been a long-standing research problem, since the rise of AI.

The work of McCarthy *et al.* (1960) was the earliest to emphasize the importance of reasoning about actions. They developed an advice-taker system that can do deductive reasoning about scenarios such as "going to the airport from home" requires "walking to the car" and "driving the car to the airport". Since then, many real-life use cases have been identified which require AI models to understand interactions among the current states of the world, actions being performed over various objects present, and most likely following states.

In a recent tweet (LeCun, 2022), Prof. Yann LeCunn has also emphasized the importance of this research direction. He mentions that "while we progress towards

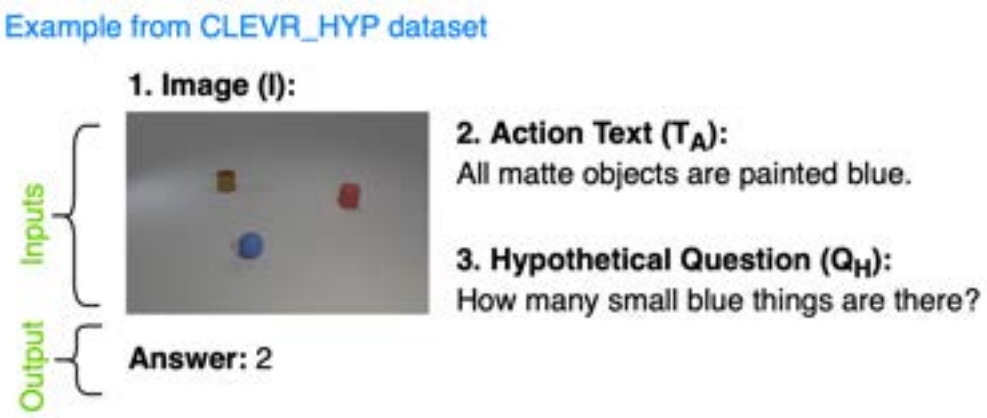


Figure 4.1: Revisiting the CLEVR_HYP Task (Sampat *et al.*, 2021): Answer a Reasoning Question (Q_H) About Changes Caused Over the Given Image (I) by Performing a Hypothetical Action (T_A)

human-level AI, I believe we need to find new concepts that would allow machines to learn to predict how one can influence the world through taking actions". While RAC has been more popular among knowledge representation and logic communities, it has recently piqued the interest of researchers in NLP and vision domains. Specifically, the work of Tandon *et al.* (2019); Wagner *et al.* (2018); Sampat *et al.* (2022b) studies whether neural models could reason about actions and changes, provided a dataset with linguistic and/or visual inputs.

This is a follow-up work on the hypothetical action reasoning task discussed in Chapter 3. To recap, a VQA task with hypothetical actions to be performed over CLEVR-like images was proposed and a synthetic CLEVR_HYP dataset (Sampat *et al.*, 2021) was proposed. An example from this dataset is shown in Figure 4.1. The key objective is to understand changes caused over a visual scene by an action described in the natural language and answer a reasoning question.

Best performing baseline models over this dataset (discussed in Section 3.4) divide the task into two phases- (i) given an image and an action, predict how the contents

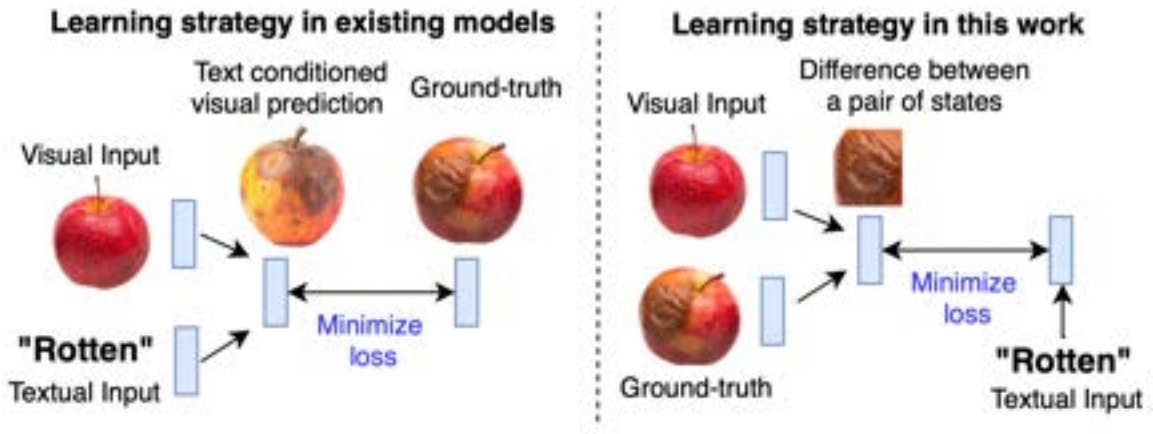


Figure 4.2: Intuition About the Learning Strategy Leveraged in This Work (Right), in Comparison With the Learning Strategy of Existing Models (Left) for Action-Effect Reasoning (Blue Box Denotes Vector Representation)

of the image will change as a result of performing an action over it and, (ii) leverage the result of phase (i) to answer a given question. The aforementioned way of solving the hypothetical reasoning task is identical to the learning strategy shown in Figure 4.2(left). Particularly, by observing visual features from an input image (i.e. features from the image of an apple) and textual representation of actions (i.e. embedding corresponding to the text 'rotten'), models attempt to imagine the possible effects (i.e. how a rotten apple would look like). The model's learning process in this case is driven by the ground-truth labels that are part of the data used for supervision.

Though this seems a viable approach from a modeling perspective, such an action-effect representation learning strategy is quite data-intensive in nature. With more variety of attributes and possible values of those attributes, the number of action-effect combinations grows exponentially. For example, in the CLEVR_HYP dataset, there were four object attributes and fifteen possible values of those attributes (8 colors, 3 shapes, 2 sizes, and 2 materials). Hence, there are 96 object attribute combinations that can be paired with each action type. Even with such a limited

synthetic domain, it is not possible to cover all variations in the training data. This indicates that models require robust compositional generalization ability to succeed on this task.

In this work (Sampat *et al.*, 2022a), an alternative learning strategy is proposed, as shown in Figure 4.2(right). This is also a supervised learning method, where the model first learns a representation corresponding to the difference of a pair of states (before and after the action is performed). In the example in Figure 4.2(right), the model first compares visual representations of a good apple and a rotten apple to understand their difference. Then, the model attempts to minimize the distance between learned visual differences between the states (i.e. the rotten portion of the apple) and the linguistic action that causes visual differences (i.e. representation of the word 'rotten'). In particular, an encoder-decoder architecture is set up to implement this learning strategy. The trained encoder-decoder is then combined with an existing visual scene parser and a scene graph question answering module to evaluate its effectiveness on the CLEVR_HYP dataset. The key contributions of this work are as follows;

- A novel learning strategy for predicting the effects of actions in the vision-language domain is proposed (as shown in the right part of Figure 4.2).
- A three-stage model is developed that implements the proposed learning strategy, and it is evaluated on the CLEVR_HYP dataset (introduced in Chapter 3).
- Through ablations and analysis, the effectiveness of the proposed model is demonstrated in terms of performance (5.9% accuracy improvements), data efficiency (one-third of training data required), and generalization capability in comparison with the best existing baselines (discussed in Section 3.4).

4.2.1 Proposed Model: Action Representation Learner (ARL)

When developing a model over the CLEVR_HYP dataset, the key focus should be on learning a mapping between actions performed and visual changes caused over the objects in the image. Figure 4.2(right) graphically demonstrates the intuition behind how such learning can be modeled. The underlying hypothesis is that a model would be able to better learn action representations by observing the difference between a pair of states (before and after the action is performed). Once the ability to robustly predict the visual differences (in a vectorized form) is achieved, another neural network can be trained to learn a representation of action (that is described in natural language) and minimize the loss between the two. In this regard, a three-stage model is proposed, as shown in Figure 4.3 (training phase) and Figure 4.5 (testing phase), which has the potential to better capture the causal structure of this task. A detailed description of each stage is provided below and can be visualized in Figure 4.4.

Stage-1 Actions have the power to change the state of the world. In other words, the difference between a pair of states can be considered as a function of performing actions. For example, consider two states 'green cube' and 'red cube'. The change between the above states is 'green $\rightarrow$ red'. A function 'paint' or 'change color' can be defined, which can simulate the desired effect i.e. 'paint(green) $\rightarrow$ red'. This stage aims to capture changes between a pair of states.

Specifically, an encoder-decoder model is set up to achieve this objective. For better understanding, various symbols are color-coded in the following text. The red symbols denote entities predicted by the component that is learned/trained in

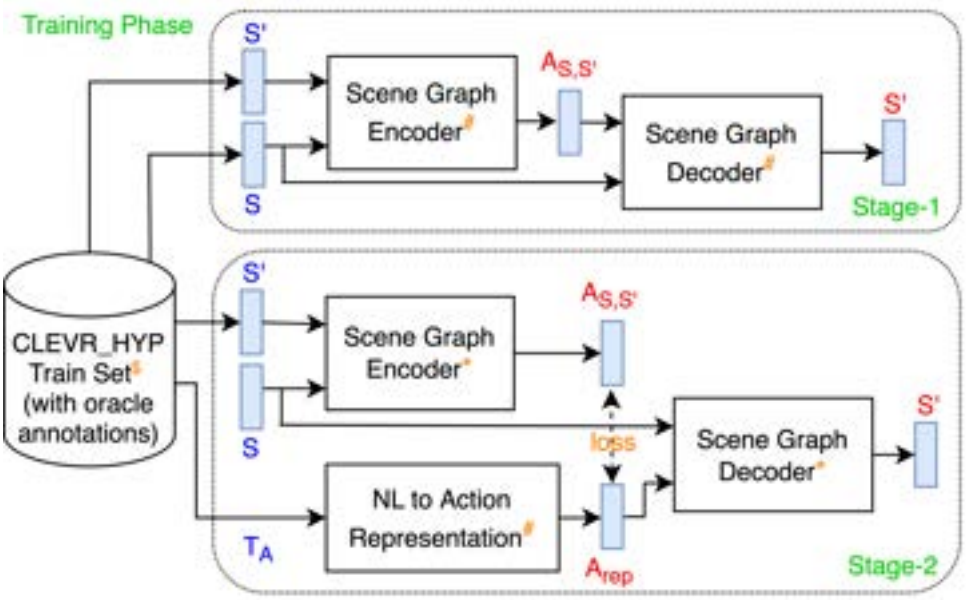


Figure 4.3: The Training Phase of the Proposed 3-stage ARL Model (Best Viewed in Color; Notations Used: # Denotes Module Being Trained in the Current Stage, * Refers to Modules Trained in Previous Stages and Frozen, $ Indicates Data Balanced by Action Types)

the proposed model. Whereas blue symbols denote entities utilized directly from the dataset.

As described in Section 3.3.1, the training set of CLEVR_HYP provides oracle annotations for an initial scene graph S (using which the image is rendered) and a scene graph after executing the action text S′. A random subset of 20k scene graph pairs from the CLEVR_HYP training set is sampled to train this encoder-decoder, which is balanced by possible action types (add, remove, change, move). Experimentation with different data sizes is performed in this work (and discussed in Section 4.3.3), however, the best results are obtained when the model is trained with 20k samples and the length of the learned action vector is 125.

A representation corresponding to the difference between states S and S′ i.e. $A_{S,S'}$

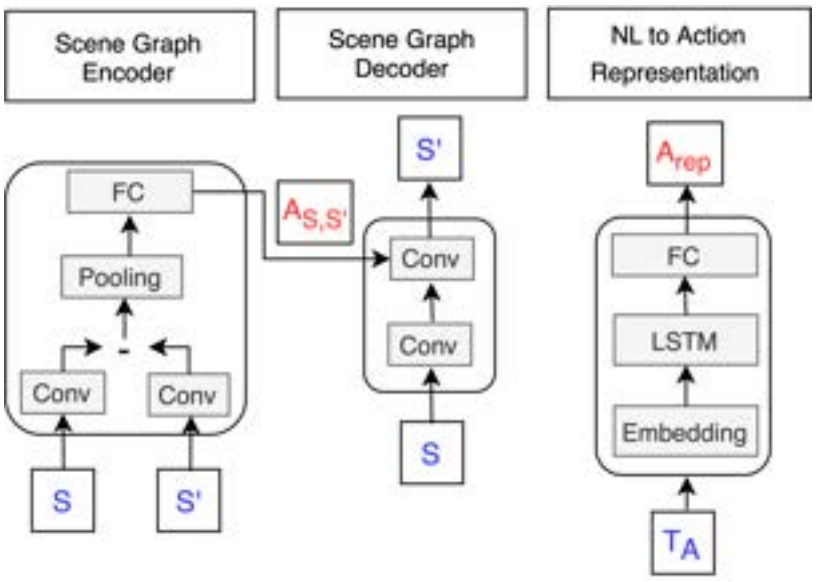


Figure 4.4: Internal Components of the Modules Trained in Stage-1 and Stage-2 (Best Viewed in Color)

is learned using the encoder. At the test time, the updated scene graph S' is not available. To address this issue, the encoder is followed by a decoder, which can reconstruct S' provided S and the learned representation of the scene difference $A_{S,S'}$ in the encoder network. Formally,

$$A_{S,S'} = SceneGraphEncoder(S,\ S') \tag{4.1}$$

$$S' = SceneGraphDecoder(S,\ A_{S,S'}) \tag{4.2}$$

The encoder-decoder networks are jointly trained with the following objective;

$$argmax_{\Theta_{Encoder}\Theta_{Decoder}}[\, logP\,(S' \mid \underbrace{Decoder(S,\ \overbrace{Encoder(S,\ S')}^{A_{S,S'}})}_{S'})\,)\,] \tag{4.3}$$

Note that, in common applications involving encoder-decoder architecture, the

decoder part of the model is removed once the desired performance is achieved and the encoder is used to encode input sequences to a fixed-length vector at test-time. Contrary in this work, the encoder part is discarded and the decoder part is retained so that at the test time updated scene representation can be predicted from an initial scene and a learned action vector.

Stage-2 The objective of this stage is to learn a network 'NL to Action Representation' (shown in stage-2 of Figure 4.3) that can map the natural language action text T_A (input provided for each CLEVR_HYP sample) to the learned action vector $\mathrm{A}_{S,S'}$. During the training of this module, encoder and decoder modules from stage-1 are frozen and used for inference. Specifically, the frozen encoder is provided with a pair of scene graphs S and S′, which will predict $\mathrm{A}_{S,S'}$. The 'NL to Action Representation' (NL2AR) network takes T_A as an input and uses encoder-predicted $\mathrm{A}_{S,S'}$ as supervision to learn A_{rep}. In other words, an optimally trained NL2AR module will convert T_A into A_{rep} such that the loss between A_{rep} and the encoder-predicted $\mathrm{A}_{S,S'}$ is minimized. Finally, A_{rep} predicted by NL2AR along with the input scene-graph S is fed to the frozen decoder to predict S′. The overall training objective for stage-2 is as follows;

$$argmax_{\theta_{NLtoAR}} [\, logP\, (\, S' \mid \underbrace{Decoder(\, S,\, \overbrace{NLtoAR(T_A)}^{\mathrm{A}_{rep}}\,)}_{\mathrm{S}'}\,)\,] \tag{4.4}$$

The NL2AR module uses a long-short term memory (LSTM) encoder in particular. The embedding object is a preliminary layer in front of an LSTM recurrent layer and a dense layer. During model training, in addition to finding the values for the weights of the LSTM and dense layers, the word embeddings for each word in the training set are computed. This is achieved using nn.Embedding(vocabulary_size,

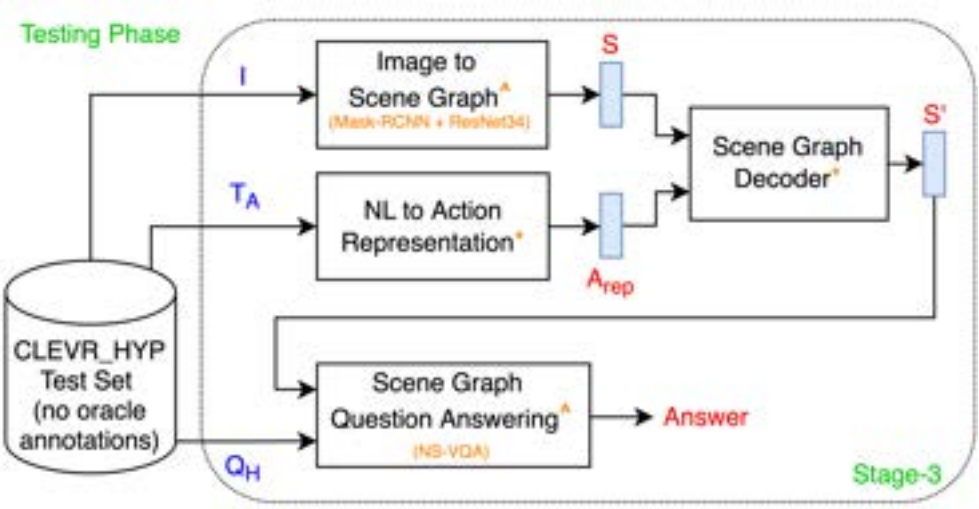


Figure 4.5: The Testing Phase of the Proposed Three-stage ARL Model (Best Viewed in Color; Notations Used: * Refers to Modules Trained in Previous Stages and Frozen, $\wedge$ Indicates Modules Adapted From Existing Works)

embedding_size) layer defined in pytorch. This way, a fixed length one-hot vector of a given length is generated for each word in the vocabulary depending on the position of the word in context and updated using back-propagation. The embedding layer is similar to a linear layer, which returns the index where one is located instead of returning the whole one-hot vector. It takes an action text T_A as a sequence of learned word embeddings, runs an LSTM over them, and then projects from the final cell state to get the output vector. The LSTM has a hidden layer of size 200.

Stage-3 Finally, in this stage, a decoder module (trained in stage-1) and NL2AR module (trained in stage-2) are combined with off-the-shelf 'image to scene-graph generator' and 'scene-graph question answering' modules to create a test pipeline with respect to CLEVR_HYP data. This workflow can be visualized in Figure 4.5 and there is no training involved in this stage. For each image in the test data, Mask R-CNN (He *et al.*, 2017) is used to generate segment proposals of all objects.

Along with generating the segmentation mask, the network also classifies the objects based on their visual attributes- color, material, size, and shape. The threshold for segment proposals is set to 0.9 i.e. segments with bounding-box scores less than 0.9 are dropped. The segment for each object, paired with the original image (resized to 224x224) is sent to ResNet-34 (He *et al.*, 2016) to extract 3D coordinates of objects in the scene. The inclusion of the original full image in ResNet-34 addition to object segments has been observed to improve performance as reported by Yi *et al.* (2018). As a result, a graph S is obtained corresponding to the input image I which captures object attributes and their spatial positions.

Then, the action text T_A for the given test instance is provided as an input to the NL2AR module trained in stage-2 and its vector representation A_{rep} is obtained. Along with the scene graph obtained at the test time (S), A_{rep} is provided to the decoder module trained in stage-1 to obtain a prediction of the updated scene graph S′. This will represent the scene graph corresponding to the test image after the action is performed over it i.e. capturing the changes in object attributes/location as a result of performing the action. Finally, the neural-symbolic scene graph question answering module developed by Yi *et al.* (2018) over CLEVR dataset (Johnson *et al.*, 2017a) is used to predict an answer to the question Q_H with respect to the updated scene graph S′ predicted by the decoder.

4.3 Results and Discussion

After the two-stage training process is complete (as per Figure 4.3), the evaluation on CLEVR_HYP (Sampat *et al.*, 2021) test data takes place as per the workflow described in Figure 4.5. The task of predicting an Answer (a) provided Image (I), Action Text (T_A) and Question (Q_H) in CLEVR_HYP is formulated as a 27-way classification. The set of answers A comprises of object attributes (8 colors + 3 shapes

+ 2 sizes + 2 material), numerals (0-9), and booleans (yes/no). Each image-action-question context has exactly one correct answer therefore, the standard accuracy metric is used for the evaluation.

4.3.1 Quantitative Results

Test performance on CLEVR_HYP			
Test Set Type / Model	TIE	SGU	ARL
Ordinary	64.7	70.5	76.4
2HopA$_T$	55.6	64.4	69.2
2HopQ$_H$	58.7	66.5	70.7

Table 4.1: Performance Comparison of Three-stage Model (ARL) Proposed in This Work With Two Baseline Models (TIE and SGU) Described in Section 3.4 Over Three Test Sets of CLEVR_HYP

The CLEVR_HYP dataset has three test sets- Ordinary, 2HopT$_A$, and 2HopQ$_H$. The ordinary test set consists of examples that require models to tackle similar complexity of actions and reasoning capabilities observed during the training. Whereas 2HopActionText (2HopT$_A$) and 2HopQuestion (2HopQ$_H$) sets are provided to test the generalization capability of the trained models. Both the generalization test sets are relatively small (only 1500 image-action text-question samples) compared to the ordinary test set (13.5k image-action text-question samples). In the 2HopT$_A$ test, the model requires to perform two different actions one after the other on the scene. For example, 'Add a small blue metal cylinder to the right of the large yellow cube and remove the large cylinder from the scene.' and 'Move the purple object on top of the small red cube then change its color to cyan.'. In the 2HopQ$_H$ test, the model requires correctly answering questions involving logical operations 'and', 'or' and 'not'. For

example, 'How many objects are either red or cylinder?' and 'Are there any rubber cubes that are not green?'.

Table 4.1 demonstrates the performance of the proposed three-stage ARL model in comparison with the two best-performing models (TIE and SGU) described in Section 3.4. The ARL model outperforms TIE and SGU baselines by 5.9%, 4.8%, and 4.2% on Ordinary, 2HopT$_A$, and 2HopQ$_H$ test sets respectively. This demonstrates that the proposed model not only achieves better overall accuracy but also has improved generalization capability when multiple actions have to be performed on the image or understand logical combinations of attributes while performing reasoning.

Further analysis is performed to gauge the model's ability to perform a particular action. The model is expected to perform one of the four actions- add objects, remove objects, change attributes, or move objects. For all three models, their performance on the validation set corresponding to each action type is shown in the upper half of Table 4.2. The results suggest that the proposed ARL method achieves better fine-grained accuracy for all action types compared to existing models. Overall, all three models do quite well on 'remove' and 'change' actions whereas struggle when new objects are added or existing objects are moved around. Yet, the ARL model shows 4.4% and 3.2% improvements on 'add' and 'move' actions respectively compared to the best previous baseline SGU.

For the 2HopT$_A$ test set, the model is expected to perform two different actions (among add, remove, change, and move) one after the other. The observation from the validation results remains consistent (as per the lower half of Table 4.2) when multiple actions are combined. In other words, models can achieve relatively high accuracy for actions 'remove' and 'change', hence their combination 'remove+change' also has a higher success rate. Whereas other combinations of actions accomplish relatively lower performance. It leads to the conclusion that understanding the ef-

Accuracy(%) by Action Types			
Validation	TIE	SGU	ARL
Add	58.2	65.9	70.3
Remove	89.4	88.6	94.1
Change	88.7	91.2	95.8
Move	61.5	69.4	72.6
$2HopT_A$	*TIE*	*SGU*	ARL
Add + Remove	53.6	63.2	66.7
Add + Change	55.4	64.7	70.6
Add + Move	49.7	57.5	63.2
Remove + Change	82.1	85.5	91.6
Remove + Move	52.6	66.4	68.3
Change + Move	53.8	63.3	67.1

Table 4.2: Performance Breakdown of Models by Different Action Types for Validation Set and 2HopT_A Test Set

fect of different actions is of varying complexity. It also indicates that the learned action representations in the proposed model ARL are helpful as they show better generalization by action types.

Understanding changes caused by actions is a core challenge in the CLEVR_HYP (Sampat *et al.*, 2021) task, however, it is formulated as a question answering downstream task. In other words, to demonstrate that a model can correctly understand the effects of actions in the CLEVR_HYP setting, the model has to correctly answer a variety of visual reasoning questions. In particular, models should be able to perform counting, check the existence of objects given the criteria, compare sets

Accuracy(%) by Reasoning Types			
Validation	TIE	SGU	ARL
Count	60.2	74.3	78.6
Exist	69.6	72.6	77.3
CompareInteger	56.7	67.3	70.7
CompareAttribute	68.7	70.5	73.4
QueryAttribute	65.4	68.1	74.9
2HopQ$_H$	*TIE*	*SGU*	*ARL*
And	59.2	67.1	70.3
Or	58.8	67.4	71.5
Not	58.1	65.0	68.4

Table 4.3: Performance Breakdown of Models by Different Question Types for Validation Set and 2HopQ$_H$ Test Set

of objects, or retrieve attributes of the desired objects to answer questions in this dataset. Therefore, a fine-grained analysis of the model performance based on their capability to perform the above-mentioned kinds of reasoning is conducted. The results are summarized in Table 4.3. For the validation set, the proposed ARL method has decent fine-grained accuracy across all reasoning types, but improvements on the 'query attribute' and 'exist' types are maximum (6.8% and 4.7% respectively). For the 2HopQ$_H$ test set, the model is expected to understand logical connectives (and, or, not). Though there is minor performance gain in this case as well, the proposed pipeline is limited by the capabilities of the question answering model of Yi *et al.* (2018) that is used in stage-3.

4.3.2 Qualitative Results

In Figure 4.6, nine qualitative examples of the scene graphs predicted by the proposed ARL model (i.e. S' in stage-3) over a variety of action texts are visually demonstrated. Examples 1-6 indicate that the model can correctly identify a combination of object attributes (color, size, shape, material) as per the given criteria. Specifically, the model is able to correctly distinguish between actions 'remove brown objects', 'remove small brown objects', and 'remove brown cylinders'. Examples 4 and 6 demonstrate that the model is consistent in predictions when synonyms of various words are used (e.g. sphere∼ball, shiny∼metallic) in the dataset. Finally, examples 7-9 show that the model does reasonably well on other kinds of actions (add, change, and move).

Further, a t-SNE (t-Distributed Stochastic Neighbor Embedding) plot of action vectors learned by the ARL model is generated as shown in Figure 4.7. Since the training data size is quite large, a random subset of 3000 samples covering each action class (add, remove, change, move in-plane, and move out-of-plane) is used to generate this plot. Also, the learned action vector in the ARL model is 125-dimensional, therefore dimensionality reduction is performed to generate this visualization. In particular, Principal Component Analysis (PCA) with n=50 is used for this purpose and then t-SNE with n=2 is used to create a 2D visualization of the learned action vectors (where n represents the dimension of the embedded space). Internally, t-SNE converts similarities between data points to calculate joint probabilities and minimizes the Kullback-Leibler divergence between the computed joint probabilities of the low-dimensional embedding and high-dimensional original data.

At first glance at Figure 4.7, it appears that the learned action representations formulate well-defined and separable clusters for each action type. Clusters for add,

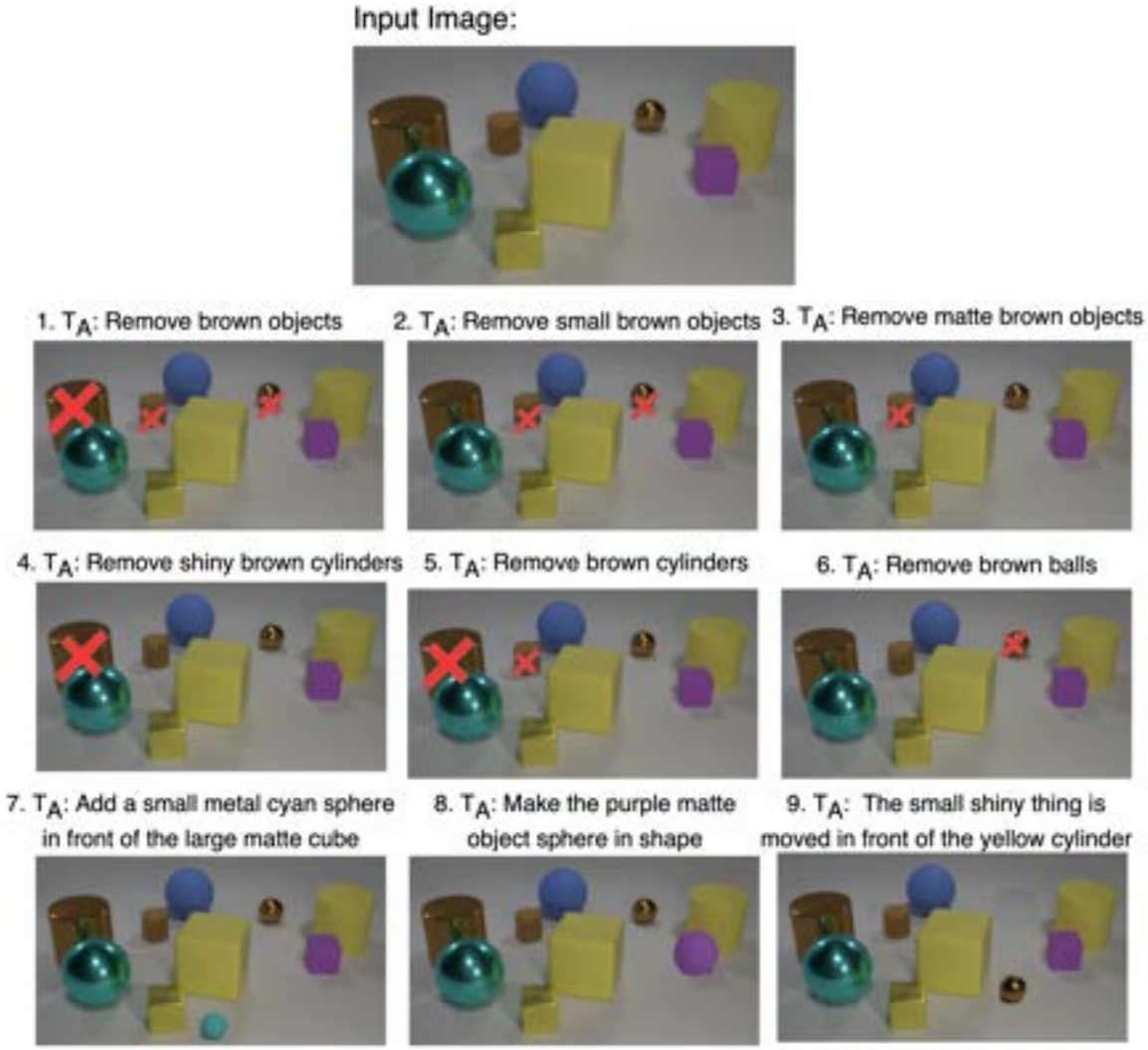


Figure 4.6: Correct Scene Graph Predictions Made by the ARL Model for Given (Input Image, Action Text) Tuples: (i) Examples 1-6 Indicate That the Model can Correctly Identify Combinations of Object Attributes Provided in the Action Text, (ii) Examples 4 and 6 Demonstrate That the ARL System is Consistent in Predictions When Synonyms of Various Words are Used in the Dataset, (iii) Examples 7-9 Show That the Proposed Model can Perform Other Actions Reasonably Well

remove and change actions are closer and somewhat overlapping. Upon further analysis, it is observed that many samples involving the 'change' action are interpreted by the reasoner as the 'remove+add' action. For example, if the color of the 'small blue metal sphere' is changed to 'red', the action reasoner interprets it as the removal of the 'small blue metal sphere' followed by the addition of a 'small red metal sphere' on the same location. This is technically correct and leads to the same desired effect of the action.

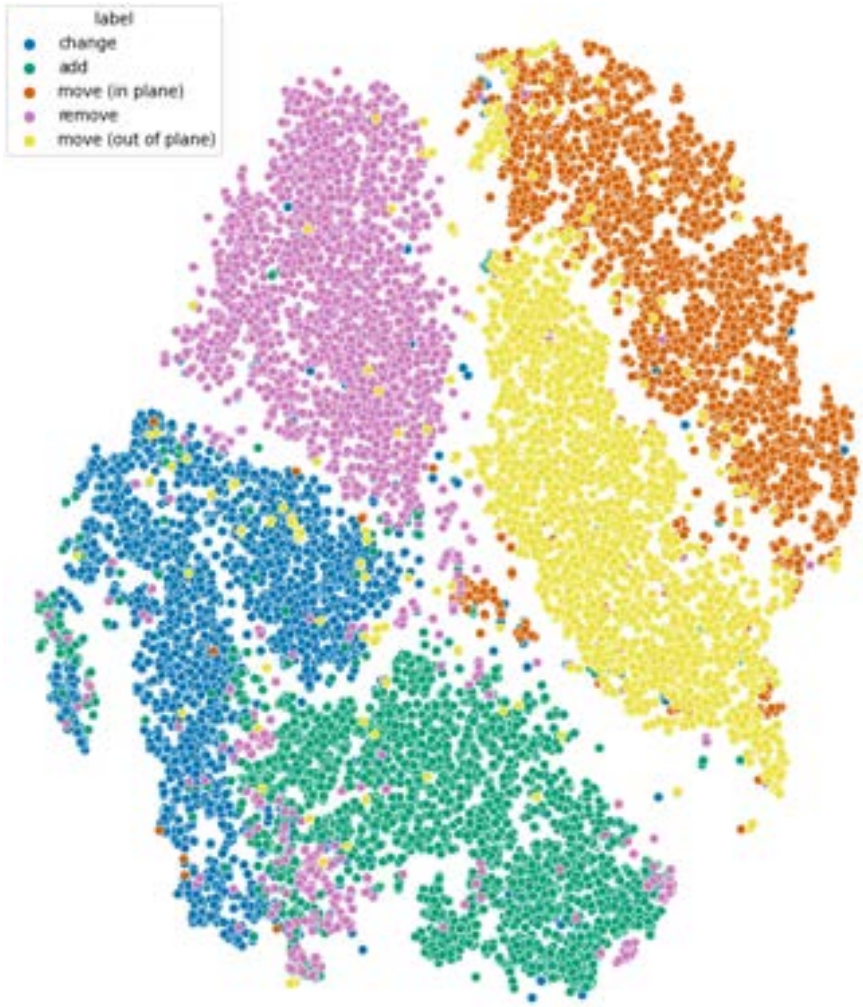


Figure 4.7: The t-SNE Plot of Learned Action Vectors

4.3.3 Ablation Study

In this section, three ablation studies are performed to demonstrate the effectiveness of the ARL model and some design choices that lead to its best performance.

Importance of stage-1 training Cause-effect learning with respect to actions is a key focus in CLEVR_HYP. The stage-1 module (described in Section 4.2.1) plays a critical role in the ARL model which is responsible for learning the causal structure of the world. To demonstrate this, two experiments are set up; first, the training takes place in a sequential manner (stage-1 followed by stage-2), where trained encoder-decoders from stage-1 are frozen and utilized in stage-2. The second experiment only consists of stage-2 training where the encoder-decoder is randomly initialized.

Task	**Experiment**	**Accuracy (%)**
Scene Graph Update	Stage(2 only)	56.3
	Stage(1+2)	87.2
Question Answering	Stage(2+3 only)	45.7
	Stage(1+2+3)	76.4

Table 4.4: Performance of the ARL Model in the Absence and Presence of Stage-1 Over Ordinary Test Set

The results corresponding to the above two experimental setups are reported in the upper part of Table 4.4. The accuracy of the setup that involves only stage-2 training is 56.3%, whereas the setup involving both stage(1+2) training achieves 87.2% performance. This indicates that the inclusion of stage-1 is effective which boosts the scene graph update accuracy by $\sim$30%.

To evaluate the question answering task of CLEVR_HYP, both the setups described above are followed by stage-3 where the image parser and scene graph question answering modules are combined to predict the answer. The results for this ablation are reported in the lower part of Table 4.4. It can be observed that the pipeline that consists of all three stages achieves 76.4% accuracy, which is significantly improved in comparison with the pipeline involving only stage(2+3).

Provided the synthetic nature of images and a limited number of object attributes in CLEVR_HYP, the image to scene graph generation is pretty accurate. Also, it is known that the scene graph question answering module used in this work (Yi *et al.*, 2018) has near-perfect performance on CLEVR (Johnson *et al.*, 2017a) dataset. As a result, the gains achieved in the scene graph update task (due to stage-1) directly benefit the question answering performance without a significant loss.

Performance with varying amounts of training data in stage-1 In this ablation, the goal is to observe the ARL model's performance when stage-1 training is done with respect to different data sizes. Particularly, stage-1 training is carried out over the dataset consisting of 1000 to 25000 samples (pair of scene graphs) in an increment of 5000. The respective results for both scene graph update and downstream question answering tasks are demonstrated in Figure 4.8(top). As observed from the graph, initially, the model is able to learn better action representations when more data is provided. However, the performance for the scene graph update and overall question answering tasks saturates to ∼86% and ∼75% respectively when training data contains more than 20k samples.

Performance with different lengths of learned action vector in stage-1 In this ablation, the goal is to find out the optimal length of action vectors that can

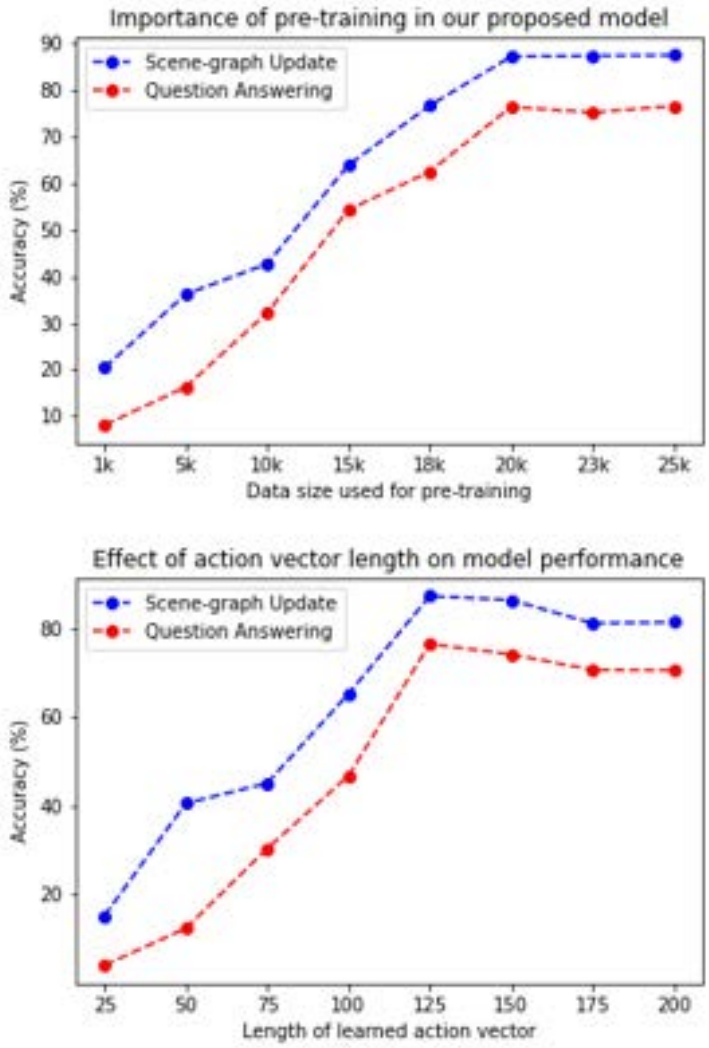


Figure 4.8: Performance of the Proposed ARL Model With Varied (Top) Data Size and (Bottom) Action Vector Lengths

reasonably capture the complexity of actions and simulate their effects. Experiment with different lengths of learned action vector- from 25 to 200 in increments of 25 is carried out and reported in Figure 4.8(bottom). As observed from the graph, initially, the model demonstrates steep accuracy improvements when the vector length is increased, with the exception of vector length 75 for the scene graph update task. However, performance reaches a peak for both scene graph update and downstream

question answering tasks when the action vector length is set to 125.

In summary, several tasks in the vision and language domain require an understanding of the causal structure of the world. In this work, a novel ARL model is proposed that can effectively learn action representations and tackles the what-if vision-language reasoning task of CLEVR_HYP dataset (Sampat *et al.*, 2021). The proposed ARL model leverages a two-step training process to model this causal task. It outperforms existing baselines while being data-efficient and shows better generalization capability on this dataset. Through qualitative results, insights on the learned action representations are provided and ablations are conducted to validate the effectiveness of the proposed method. Extension of this approach to a larger set of actions would enable the possibility of AI agents that are equipped with action-effect reasoning capability and can better collaborate with humans in the physical world.

Chapter 5

VL-GLUE: A SUITE OF FUNDAMENTAL YET CHALLENGING VISUO-LINGUISTIC REASONING TASK

5.1 Motivation

Deriving inference from multi-modal artifacts is an important skill for humans to navigate day-to-day situations. For example, reading through product manuals/user guides, driving, and understanding content from textbooks and other documents (including newspapers and recipes) that are rich in visual and textual content. In the above scenarios, the language modality provides important cues for decision-making or aids in grasping novel concepts along with the visual information, and not only serves as an evaluation mechanism (unlike most existing computer vision tasks).

To train vision-language systems that can perform multi-modal reasoning, many tasks have been proposed (Reddy *et al.*, 2022; Talmor *et al.*, 2020; Chen *et al.*, 2020b; Sampat *et al.*, 2020a, 2021; Singh *et al.*, 2021). Though state-of-the-art AI systems demonstrate accuracy at par with human levels on a variety of vision and language tasks in isolation, their performance over aforementioned benchmarks is remarkably low.

One underlying possibility for this performance gap is due to models being over-reliant on image-text similarity (during the pre-training phase) to solve various V&L tasks and not actually learning to combine information from visual+textual modalities. Moreover, research efforts in this direction have been limited due to two challenges concerning the above benchmarks: (i) relatively smaller training data available (compared to other V&L tasks such as VQA or image captioning), and (ii) hetero-

geneous task formats across datasets, i.e. some datasets have multiple choice QA whereas others have open-ended/generative QA or retrieval (from a set of candidate documents) followed by QA, etc.

To this end, in this work, GLUE-like (Wang *et al.*, 2019b, 2018) benchmark for Visuo-Linguistic understanding (VL-GLUE) is proposed. VL-GLUE is a multi-task benchmark consisting of seven different tasks that at their core test for visual-linguistic reasoning skills of models. There is a hierarchy of VQA tasks depending on the complexity of the visuals incorporated, the reasoning abilities required, and the requirement of necessary knowledge to answer questions. The proposed benchmark combines all such variations and consists of 106k problems spread across seven different tasks.

The motivation behind VL-GLUE is similar to GLUE (Wang *et al.*, 2019b, 2018) and NUMGLUE (Mishra *et al.*, 2022), which are multi-task benchmarks aimed at the development of models that can demonstrate superior language understanding and mathematical reasoning abilities respectively. Different from these works, VL-GLUE is designed to progress toward AI systems that are capable of performing visuo-linguistic reasoning in a general setting. Achieving superior performance on this benchmark would require models' ability to perform joint reasoning over provided visual and linguistic inputs without relying on task or dataset-specific signals.

In summary, there are three key contributions of this work:

- A brief survey of existing GLUE-like benchmarks is conducted, which are precursors to this work.
- VL-GLUE, a novel multi-task benchmark consisting of seven different tasks is proposed, solving which requires an ability to make inferences by combining a portion of visual and textual information provided as a context.

- It is demonstrated that VL-GLUE is a challenging benchmark for large-scale vision-language models, obtaining poor scores not only in zero-shot settings but also after fine-tuning. While there are plenty of visuo-linguistic applications, this fundamental barrier needs to be addressed with utmost priority.

5.2 A Brief Survey of GLUE-like Benchmarks

Towards the goal of creating a better general-purpose language technology (rather than catering to individual models specific to the given domain or use case), Wang *et al.* (2018) developed the GLUE benchmark. There are three key characteristics of this benchmark: (i) inclusion of more than one linguistic task (including a mix of sentiment analysis, textual entailment, and sentence similarity), (ii) varied size and genres of training data to facilitate sample-efficient learning yet encouraging effective knowledge-transfer across tasks, and (iii) built upon preexisting datasets that are challenging and interesting, as agreed upon by the researchers. Since its proposal, GLUE (Wang *et al.*, 2018) has become a prominent evaluation framework for research towards general-purpose language understanding technologies.

Following their characteristics, there have been several efforts to create similar benchmarks for non-English languages, broader NLP tasks (such as text generation, dialog understanding, and arithmetic), modalities beyond language (speech, image, and videos), and other learning paradigms (such as few-shot understanding and out-of-distribution robustness). A brief survey of GLUE-like benchmarks is summarized in Table 5.1 and 5.2. For each benchmark, their focus area (within the scope of NLP research), modalities present, and number of diverse tasks incorporated in the benchmark are listed.

Benchmark Name	Focus Area	Modality	#Tasks
GLUE (Wang *et al.*, 2018)	English language understanding	Text	9
Super-GLUE (Wang *et al.*, 2019a)	English language understanding	Text	8
FewGLUE (Schick and Schütze, 2021)	Few-shot English language understanding	Text	8
CLUES (Mukherjee *et al.*, 2021)	Few-shot English language understanding	Text	6
KLEJ (Rybak *et al.*, 2020)	Polish language understanding	Text	9
GLUES (Cañete *et al.*, 2020)	Spanish language understanding	Text	7
SwedishGLUE (Adesam *et al.*, 2020)	Swedish language understanding	Text	10
CLUE (Xu *et al.*, 2020)	Chinese language understanding	Text	9
FewCLUE (Xu *et al.*, 2021)	Few-shot Chinese language understanding	Text	9
RussianSuperGLUE (Shavrina *et al.*, 2020)	Russian language understanding	Text	9
ALUE (Seelawi *et al.*, 2021)	Arabic language understanding	Text	8
FLUE (Le *et al.*, 2020)	French language understanding	Text	6
JGLUE (Kurihara *et al.*, 2022)	Japanese language understanding	Text	6
KLUE (Park *et al.*, 2021)	Korean language understanding	Text	8
INDOLEM (Koto *et al.*, 2020)	Indonesian language understanding	Text	7
DialoGLUE (Mehri *et al.*, 2020)	Dialogue (English) language understanding	Text	4
AdvGLUE (Wang *et al.*, 2021)	Robustness of language models against adversarial attacks (English)	Text	14

Table 5.1: A Brief Survey of GLUE-like Benchmarks: Comparison by Focus Area (Within the Scope of NLP Research), Modalities Present and Number of Diverse Tasks Incorporated in the Benchmark. $^{\diamond}$ Indicates Multi-modal Benchmarks and † Indicates Multi-lingual Benchmarks

Benchmark Name	Focus Area	Modality	#Tasks
NUMGLUE (Mishra *et al.*, 2022)	Arithmetic understanding (English)	Text	8
GLGE (Liu *et al.*, 2021)	English language generation	Text	8
LexGLUE (Chalkidis *et al.*, 2022)	Legal language understanding (English)	Text (Legal)	7
GLUE-X (Yang *et al.*, 2023a)	Out-Of-Distribution (OOD) robustness in language understanding (English)	Text	7
XGLUE (Liang *et al.*, 2020)	Language understanding and language generation	Text†	11
IndicGLUE (Kakwani *et al.*, 2020)	Indic language understanding	Text†	6
ScandEval (Nielsen, 2023)	Scandinavian language understanding	Text†	4
XTREME (Hu *et al.*, 2020)	Cross-lingual generalization	Text†	9
GLUECoS (Khanuja *et al.*, 2020)	Code-switched language understanding	Text†	6
CodeXGLUE (Lu *et al.*, 2021)	Code understanding and code generation	Text (Code)†	10
ASR-GLUE (Feng *et al.*, 2021)	English language understanding through Automatic Speech Recognition (ASR)	Speech	6
CBLUE (Zhang *et al.*, 2022)	Biomedical language understanding (Chinese)	Text (Biomed.)	8
SLUE (Shon *et al.*, 2022)	Spoken language understanding	Speech	4
GEM-I (Su *et al.*, 2021)	Image-language understanding	Text, Image$^{\diamond\dagger}$	2
GEM-V (Su *et al.*, 2021)	Video-language understanding	Text, Video$^{\diamond\dagger}$	2
VALUE (Li *et al.*, 2021)	Video-language understanding	Text, Video$^{\diamond}$	11
VL-GLUE (this work)	Visuo-linguistic understanding	Text, Images$^{\diamond}$	7

Table 5.2: (Continued) A Brief Survey of GLUE-like Benchmarks: Comparison by Focus Area (Within the Scope of NLP Research), Modalities Present and Number of Diverse Tasks Incorporated in the Benchmark. $^{\diamond}$ Indicates Multi-modal Benchmarks and † Indicates Multi-lingual Benchmarks

5.3 VL-GLUE Benchmark Creation

The proposed VL-GLUE benchmark consists of a broad variety of visuo-linguistic (VL) reasoning tasks, which are compiled as a subset of existing datasets or their modifications. In this section, a brief overview of tasks that are part of VL-GLUE is provided. It is followed by a short description of how data items for each task are curated, and their diverse linguistic and visual nature is analyzed.

Specifically, the proposed VL-GLUE benchmark is a collection of seven different tasks that together include 106k image-passage-question-answer tuples. The tasks may either be self-contained or may require additional background knowledge (e.g. commonsense reasoning) to arrive at the final solution. However, all the tasks, at their core, involve joint reasoning over the image(s) and a passage to answer questions. Table 5.3 summarizes different task types considered in this benchmark along with the total number of data points included. It is important to note that data for each task type is collected from different sources. Depending on its source, each task may have a varied number of samples. For example, there are only 1854 examples for Task 4, whereas there are 53.4k questions under Task 6. The datasets are retained in an imbalanced manner following Wang *et al.* (2019b, 2018); Mishra *et al.* (2022).

5.3.1 Benchmark Data Collection

Task 1: VL Reasoning over Synthetic Images Synthetic or controlled dataset collection methods have been shown effective for many vision-language problems in terms of scalability, bias control in the data, and due to their inexpensive nature (in comparison with crowd-sourced alternatives). Therefore, this task encompasses data where images are synthetically rendered or collected in a controlled environment but require visual-linguistic reasoning to solve them. Images of this kind typically have a

Task	Modality Types	Size
Task 1	*Synthetic Images*	6.5k
	Includes very simple images with a few objects and limited attributes; questions test what-if reasoning and planning abilities of the model	
Task 2	*Natural Images*	2116
	Includes diverse indoor/outdoor images and questions that test complex object grounding and varied reasoning skills including commonsense	
Task 3	*Charts/Graphs*	3920
	Includes diverse chart figures (pie, bar, line plots, etc.) and templated questions that require basic math reasoning (e.g. min, max, average)	
Task 4	*Freeform Figures*	1854
	Includes visuals with complex information representation beyond standard chart types (simple template-based rendering does not work)	
Task 5	*Images+Text+Tables*	23.7k
	Includes tabular information as a part of the text context along with visuals that are useful in answering given questions	
Task 6	*Procedural Knowledge*	53.4k
	Includes diverse indoor/outdoor images; Solving this subset requires procedural knowledge of various activities such as cooking, crafts, etc.	
Task 7	*World Knowledge*	14.7k
	Includes disambiguation of various named entities (e.g. Barack Obama, white house, etc.) referred to in the text and images	

Table 5.3: A Summary of Task Types Incorporated in the Proposed VL-GLUE Benchmark

limited set of visual attributes that are fixed before the dataset is generated. Which in turn, provides a flexibility to test specific reasoning skills and possibly avoid failures due to object recognition challenges.

There are two resources used for this task, CLEVR_HYP (Sampat *et al.*, 2021) and BlocksWorld (Gokhale *et al.*, 2019). In CLEVR_HYP, a synthetic image and an action (described in natural language) are provided as inputs. The model has to perform an action over the image and visualize the effects of those actions by answering reasoning questions. CLEVR_HYP fits well under the visuo-linguistic reasoning task as without jointly reasoning over the action (natural language input) and a given image, it is not possible to correctly answer the question. A subset of CLEVR_HYP is included as a part of VL-GLUE.

On the other hand, the BlocksWorld dataset developed by Gokhale *et al.* (2019) was originally proposed to support visual planning over a pair of images. To include it under this benchmark, it was manually processed to obtain text passage and QA pairs. For example, given a pair of images (calling it a left and a right image), the passage would describe rules for moving blocks. For example, only one block can be moved at a time and a block can be placed on another block only if there is no other block on its top. Then a set of questions will be asked based on this context, such as 'How many times the red block will be moved if the left image is to be transformed into the right image?'. Figure 5.1 demonstrates two examples based on BlocksWorld and CLEVR_HYP included in the VL-GLUE benchmark.

Task 2: VL Reasoning over Natural Images While synthetic datasets have their advantages, they limit the model's understanding to a small set of objects and a constrained amount of visual attributes. This is a strict assumption considering real-life situations. Therefore, many datasets are developed with natural images that

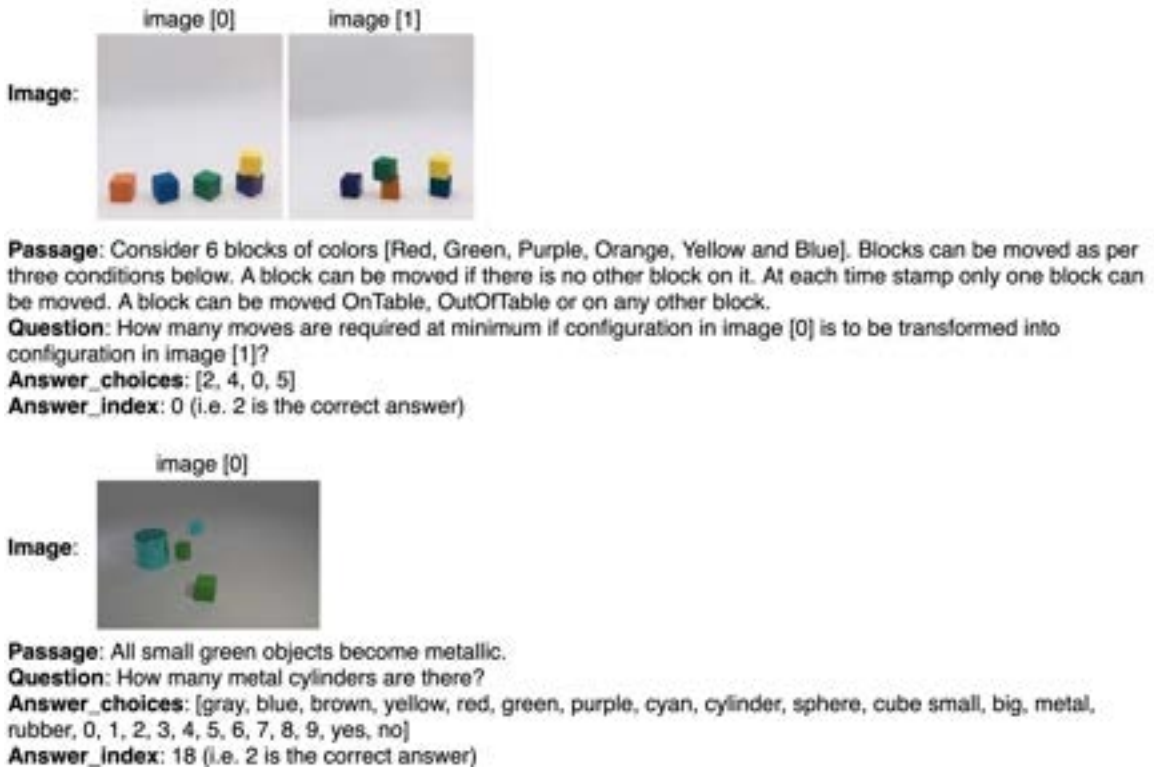


Figure 5.1: Examples From Task 1 of the VL-GLUE Benchmark (Top) BlocksWorld Dataset (Gokhale *et al.*, 2019) Repurposed as Visuo-Linguistic Reasoning Problem (Bottom) Data Items From CLEVR_HYP (Sampat *et al.*, 2021) Directly Incorporated Into the Benchmark as it is Visuo-Linguistic in its Original Form

include everyday indoor/outdoor scenes that humans encounter frequently. Among them, COCO (Chen *et al.*, 2015), NLVR (Suhr *et al.*, 2019), PIQA (Bisk *et al.*, 2020b) and WinoGround (Thrush *et al.*, 2022) datasets are leveraged to be a part of VL-GLUE as demonstrated in Figure 5.2.

Particularly, COCO, NLVR and WinoGround are existing datasets for image captioning, visual-textual classification, and image-text matching tasks respectively. The caption or sentences provided with the original datasets are converted into a passage manually. There are two kinds of questions formulated; One is a binary classification

Figure 5.2: Examples From Task 2 of the VL-GLUE Benchmark (Top & Mid) Binary Visuo-Linguistic Classification Questions Based on NLVR (Suhr *et al.*, 2019) and COCO (Chen *et al.*, 2015) Datasets Respectively (Bottom) Image Selection Visuo-Linguistic Problem Based on PIQA (Bisk *et al.*, 2020b)

question (as a True/False) over the image+passage i.e. for the given image, state whether or not the information provided in the passage is true. Second, is an image selection question, where more than one images are provided along with the passage. The goal here is to select the image that matches the description stated in the passage.

The PIQA, in its original form, is a text-only dataset for physical commonsense reasoning in the question answering form. Specifically, provided a goal that the user wants to achieve, there are two alternatives from which the model has to select one which is more plausible towards achieving that goal. To convert it into a visuo-linguistic data item, the goal text is considered as a passage. Images corresponding to the alternatives are obtained through keyword-based image crawling. The visual-

linguistic version of PIQA with crawled images and a passage (goal) is turned into image selection kind of questions described above.

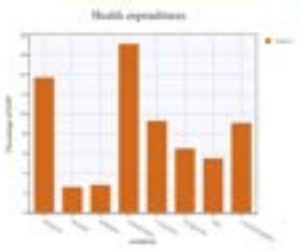
Task 3: VL Reasoning over Charts/Graphs Charts are important visual tools when large quantities of data are to be represented concisely. Also, charts are ubiquitous in various documents such as newspapers and reports and are considered to be an integral part of modern literacy. As a result, the design of standardized/psychometric tests like the GRE and PISA involves questions about chart representations. Inspired by this, this third category of VL-GLUE benchmark tests visual-linguistic understanding of AI models with respect to charts.

The data compilation for this task was based on automatic and manual efforts. Particularly, publicly available archives and test preparation materials of standardized tests like PISA and GRE were obtained. The items that involve reasoning with respect to charts and additional text were manually filtered from the test materials. There were a few challenges with this process; Many worksheets were in the form of scanned documents which propagated OCR (optical character recognition) errors when attempting to convert into digital versions. Secondly, a lot of online test materials were subject to copyrights which had to be forgone. This turned out to be quite a time-consuming and cumbersome process. As a result, there was only a small subset of chart-passage-question-answer tuples were obtained through this process.

For further scaling of data in this category, inspiration was drawn from synthetic chart QA datasets (Kahou *et al.*, 2017). To do so, tabular data from CIA 'world factbook' (Central Intelligence Agency, 2019) was obtained as a first step. This tabular data was converted into various figures like bar charts, pie charts, scatter plots, etc. For most examples, passages & questions were created from templates, answer choices were manually curated and the correct answer was annotated. Figure

image [0]

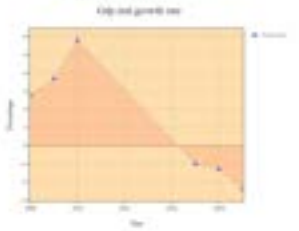
Image:



Passage: The above visualization shows the percentage distribution of GDP for health expenditures across countries. Tax revenue is an important statistical determinant towards universal health coverage and it is determined that developing countries with higher tax revenues tend to spend more on healthcare.
Question: As per the data, Which country is most likely to have the least tax revenue?
Answer_choices: [Iraq, Indonesia, Pakistan, Maldives]
Answer_index: 2 (i.e. Pakistan is the correct answer)

image [0]

Image:



Passage: GDP - real growth rate compares GDP growth on an annual basis adjusted for inflation and expressed as a percent. When the economy is expanding, the GDP growth rate is positive. But if it expands beyond 3-4%, then it could hit the peak. At that point, the bubble bursts and economic growth stalls.
Question: In which year, the economy of Puerto Rico is in danger of stalling?
Answer_choices: [2008, 2009, 2010, 2012]
Answer_index: 2 (i.e. 2010 is the correct answer)

Figure 5.3: Examples From Task 3 of the VL-GLUE Benchmark Based on CIA Factbook Data (Top) Bar Chart Demonstrating GDP% Allocated to Healthcare Expenditure for Different Countries (Bottom) Line Chart Demonstrating Puerto Rico's GDP% Over Years

5.3 demonstrates two such examples.

Task 4: VL Reasoning over Freeform Figures Beyond standard chart types, there exists a plethora of visual representations such as timelines, cycles, flowcharts, maps, etc., which do not necessarily follow a particular template like bar/line/part charts. Such visuals are also abundant in textbooks and widely used in psychometric tests like PISA. Often, such figures are more complex in comparison with images accompanied in all previous tasks, due to two reasons (as observed from the compiled data); First, they often involve multiple interrelated sub-components within the image. Second, the image incorporated in each data instance is very specific to the problem at hand and not necessarily valid/applicable to other similar problems.

Refer to Figure 5.4 for examples from the PISA test incorporated under this task. The first example requires inferring the time difference between two people living in different countries to answer the given question. Whereas, the second example demonstrates a scenario where a carpenter wants to build a shelf as shown in the image. The paragraph describes a set of resources the carpenter has, and the question asks about how many such shelves he can make at maximum provided the resources. As explained earlier, the above two examples are very specific to the scenario posed and not quite useful if the people were in different sets of countries or a different type of shelf needs to be built.

Task 5: VL Reasoning over Image+Text+Tables MultimodalQA (Talmor *et al.*, 2020) is a benchmark for reasoning over visual, textual, and tabular data constructed using Wikipedia as a source. Talmor *et al.* (2020) compiled ∼30k samples which include questions based on both unimodal and multi-modal inputs. The goal of the VL-GLUE benchmark is to be able to perform joint reasoning over image and text.

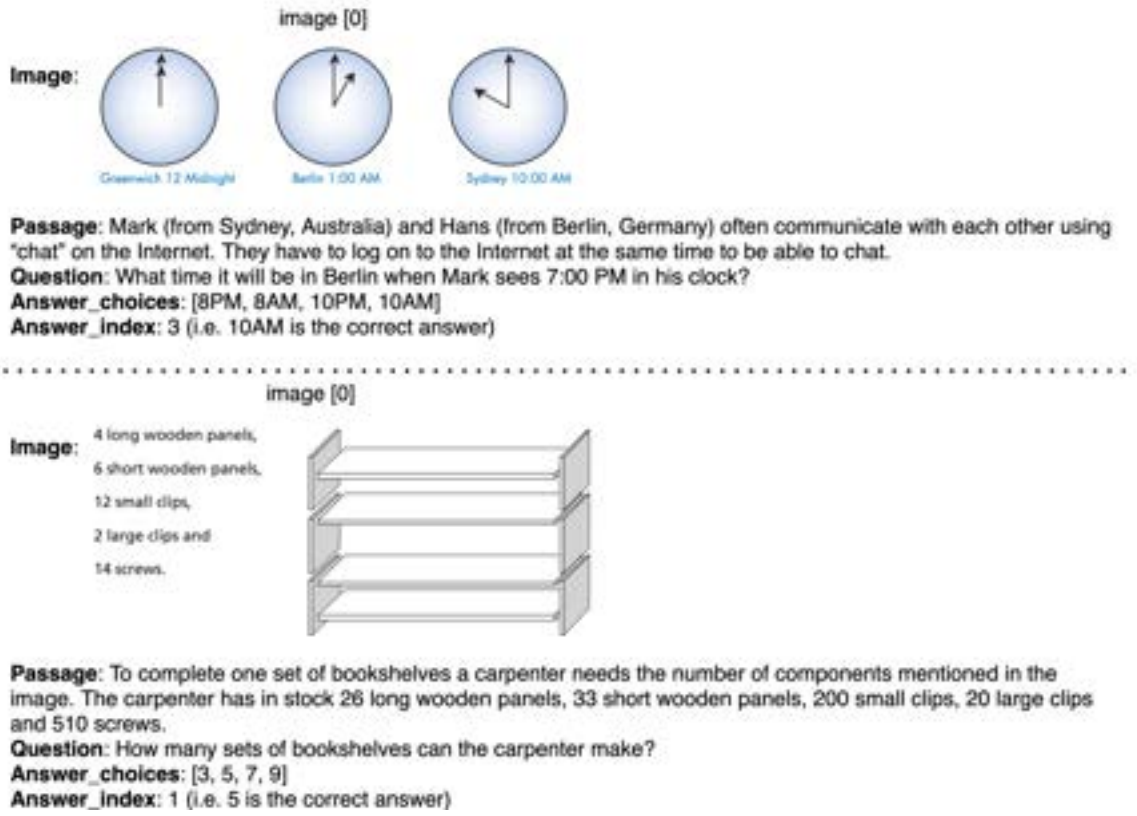


Figure 5.4: Examples From Task 4 of the VL-GLUE Benchmark Based on the PISA Tests That Involve Freeform Figures Which are Very Specific to the Problem at Hand

Therefore, a subset of the MultimodalQA dataset that requires cross-modal reasoning is filtered from the dataset and incorporated in the VL-GLUE benchmark. Using the metadata provided for each item in the MultimodalQA (about which modality among text, image, and table are necessary to answer the question), we select two subsets- (i) items that require reasoning over images and text, and (ii) items that require reasoning over images and table+text inputs.

The text and table components in MultimodalQA are quite large in most cases (which range from a couple of paragraphs to the entire Wikipedia page). The focus of VL-GLUE is on multi-modal reasoning and not on the retrieval or localization of information. Therefore, we narrow down the textual contexts by using answer span

image [0]

Image: 

Passage: The following is a list of the 20 golfers who have risen to the top of the Official World Golf Ranking. As of January 21, 2018, Dustin Johnson is the number one ranked golfer. Tiger Woods has spent the most consecutive weeks (281) and most total weeks (683) in that position. Three golfers have spent an entire calendar year atop the rankings: Nick Faldo (1993), Greg Norman (1996), and Woods (2000, 2001, 2002, 2003, 2006, 2007, 2008, 2009).
Question: When was most recent time the African-American player (in the image) in the third round of the Masters Tournament of 2010 was number 1?
Answer_choices: [2009, 2008]
Answer_index: 0 (i.e. 2009 is the correct answer)

Figure 5.5: An Example From Task 5 of the VL-GLUE Benchmark, Which is a Subset of MultimodalQA (Talmor *et al.*, 2020) Containing Image+Text Context; Without Correctly Recognizing the Person in the Image (i.e. Tiger Woods for the Above Example) and Inferring the Information Provided in the Passage (i.e. Years When Tiger Woods was a Top-ranked Golf Player), the Given Question Cannot be Correctly Answered

annotations provided in the original dataset to truncate unnecessary content that is not useful for answering a given question. In other words, the MultimodalQA dataset provides beginning and ending spans of the answers based on where they are located in the long textual input. Using these annotations, only the portion of the text that has the answer is kept.

Similarly, for many questions, more than one image is present in the respective Wikipedia pages. In this scenario, all images are merged into a single image before including them in VL-GLUE. Since this dataset was originally compiled from Wikipedia, it spans a wide variety of topics including- films, transportation, video games, industry, theatre, music, television, geography, history, literature, economy, sports, science, and politics. Therefore, the data items for Task 5 included in VL-

GLUE also reflect the diversity of topics captured by MultimodalQA. Figure 5.5 demonstrates an example from the sports category.

Task 6: VL Reasoning involving Procedural Knowledge Humans observe various actions being performed by other humans (physically or in videos/images) and over time, they build a cumulative knowledge repository of various procedural concepts. Such concepts include determining the aspects of the world that make action execution possible, predicting how the world will change as a result of the action, high-level goals associated with actions, and temporal dependency among the actions. They can effortlessly leverage this knowledge to reason in novel situations such as performing similar activities on another set of objects.

Cooking and Do-It-Yourself activities are two domains that require frequent application of such procedural concepts. To test the AI system's ability for procedural understanding, the RecipeQA (Yagcioglu *et al.*, 2018) dataset has been developed. In particular, RecipeQA is developed from cooking recipes that are multi-modal in nature. Specifically, each recipe in this dataset has a certain number of steps described both visually and textually. Hence, to convert it into a VL format, from each recipe, four steps are chosen in their temporal order. For any two steps (randomly chosen), the respective textual description is obtained. For the remaining two steps, respective images are obtained. Then the obtained textual and visual modalities are shuffled and the model is asked to arrange them in the correct temporal order as a four-way text classification problem (as shown in Figure 5.6). WikiHow (Yang *et al.*, 2021a) is used as another resource to collect data for this task. The format of wikiHow data items as step-ordering tasks over visual and textual steps is quite similar to that of RecipeQA. However, Wikihow-based dataset partition is much more diverse in terms of activities (including product assembly, gardening, crafts, etc.) and procedural knowledge in

Passage: IV. Image [0] III. Image [1] I. You will need: Cup of black coffee with sugar (how much you want, I used 2 spoons) Cup of milk Ice cube tray Blender II. Take your ice cube tray and pour the coffee in and put in the freezer till frozen.
Question: Choose the correct order of steps I-IV in order to make iced coffee frappe. Left image is denoted by [0] and right image is denoted by [1].
Answer_choices: [I-II-IV-III, I-IV-III-II, II-III-I-IV, II-I-IV-III]
Answer_index: 0 (i.e. I-II-IV-III is the correct answer)

Figure 5.6: An Example From Task 6 of the VL-GLUE Benchmark, Which Requires Procedural Knowledge of Activities That People Perform in Day-to-day Life (Such as Cooking, Gardening, Product Assembly, Arts, Personal Care, etc.)

comparison with RecipeQA.

Task 7: VL Reasoning involving World Knowledge MuMuQA (Reddy *et al.*, 2022) and WebQA (Chang *et al.*, 2022) are two large-scale multi-modal datasets compiled from Wikipedia, which is rich in terms of notable events in history, politics, and sports as well as consists of the wide variety of information about literature, culture, movies, etc. As a result, multi-modal questions in both the above datasets often include named entities (such as Barack Obama, White House, Tanabata festival, Oktoberfest, etc.) and require relevant knowledge (such as Barack Obama was a former president of the United States, Oktoberfest takes place in Germany, etc.). This aspect of both datasets poses a greater challenge in terms of recognition of entities present in the image, text understanding with named entities as well the need for relevant external knowledge. Hence, this is the most complex task category among the VL-GLUE benchmark.

The MuMuQA dataset is already visuo-linguistic in nature, therefore we readily

adapt in our benchmark without any post-processing. Following is a brief description of how the MuMuQA dataset (Reddy *et al.*, 2022) was originally constructed. Firstly, pairs of image-passage from Wikipedia were obtained which overlap in terms of entities present in the images and their mentions in the image caption and passage. Then Question-Generation Question-Answering (QGQA) models were leveraged to automatically create question-answer pairs about the image-passage context involving those named entities. Then the authors replace the span of the named entity in the question with corresponding visual attributes from the image. For example, consider an automatically generated question about a named entity [NE] for an image-passage context of the form 'What is [NE] accused of?'. The [NE] span is substituted with the referring expression 'the person wearing a yellow tie' which refers to the [NE] in the image. Hence, the final question in the MuMuQA dataset would be of the form 'What is the person wearing the yellow tie accused of?'. In this way, it is ensured that the questions are only answerable by joint reasoning over image+passage, which fits well under the objective of VL-GLUE. An example is shown in Figure 5.7.

Different from MuMuQA, the WebQA (Chang *et al.*, 2022) provides a question and a list of Wikipedia pages, which may (positive source) or may not (distractor sources) be helpful while answering the question. In other words, the objective of this dataset is to equip models with the ability to perform a careful selection of relevant sources and then perform multi-modal reasoning over the long-tailed contexts to find the answer. A similar post-processing approach to the MultimodalQA dataset (Task 5) is used to truncate unnecessary textual contexts and remove distractor images. Eventually, a subset of the WebQA image-passage pairs are manually verified to prioritize the joint image+text reasoning incorporated in this benchmark.

image [0]

Image: 

Passage: A Quarter of US States Hold Keys to Presidential Outcome \nThe U.S. has 50 states, but its national presidential election is likely to be decided in about a quarter of them.\nThey are often called battleground states, where surveys show that voters are closely split in deciding whether to give the Democratic incumbent, President Barack Obama, a second four-year term in the White House, or come January, make his Republican challenger, one-time venture capitalist Mitt Romney, the American leader. They are the states that sometimes swing from election to election in their support for Democratic or Republican candidates, whether for president or lawmakers in Congress.\nVoters across the country are now weighing their choice in advance of the November 6 election, with residents in some states already starting to cast ballots under early-voting provisions. President Barack Obama and Republican presidential candidate, former Massachusetts Gov. Mitt Romney
Question: States that are likely to vote for person in the left of the image in the presidential election
Answer_choices: ["large states", "coastal states", "red states", "battleground states"]
Answer_index: 3 (i.e. battleground statesis the correct answer)

Figure 5.7: An Example From Task 7 of the VL-GLUE Benchmark, Compiled From MuMuQA (Reddy *et al.*, 2022) Dataset Which Requires World Knowledge; The Question Refers to the Person on the Left Side of the Given Image (i.e. Barack Obama) and Asks for the States Which are Likely to Vote for Him (Which can be Found in the Passage)

5.3.2 Dataset Partitions

The data corresponding to each task is partitioned into training (80%), validation (10%), and test (10%) sets. In the cases where there are multiple questions based on the same image or passage, they are assigned to the same data partition to discourage any data leakage thereby, allowing models to potentially rely on memorization to arrive at the correct answer. All of the VL-GLUE data are in the form of classification QA ranging from 2-way answer choices to 27-way answer choices, therefore accuracy is reported as an aggregate measure of performance.

5.4 Experiments

5.4.1 Heuristic Baseline (Random Selection)

Most data items in the VL-GLUE benchmark are formulated as 4-way and 2-way multiple choice questions (MCQs) where each answer choice is likely to be picked with 25% and 50% chance respectively. The only exception, in this case, is data from CLEVR_HYP, which is formulated as a 27-class classification, for which random accuracy is 3.7%. For each task type, the accuracy of answers randomly picked for a pool of questions is computed and reported if it is correct.

5.4.2 Uni-modal Baselines

Three unimodal baselines are used for automated quality assurance of the VL-GLUE benchmark (which are not trained or fine-tuned), to prevent models from exploiting biases in the data. In particular, question-only (Q-only), passage+question only (PQ-only) and image+question only (IQ-only) models are implemented using GPT-3 (Brown *et al.*, 2020), RoBERTa (Liu *et al.*, 2019) finetuned on RACE (Lai *et al.*, 2017) dataset and BLIP (Li *et al.*, 2022a) finetuned on VQA (Antol *et al.*, 2015a) datasets respectively. The expectation here is that these unimodal baselines should ideally demonstrate lower performance on the VL-GLUE benchmark data if it indeed requires performing joint reasoning over the image and text. In other words, any system that ignores either the passage or image modality is likely to demonstrate poor performance on the VL-GLUE benchmark.

5.4.3 Multi-modal Baselines (Prediction-only)

Recently, several attempts have been made to derive transformer-based pre-trainable generic representations for visual and text modalities. Among them, top-performing

single-model architectures that support VQA tasks include BLIP (Li *et al.*, 2022a), GIT (Wang *et al.*, 2022a) and ViLT (Kim *et al.*, 2021). These three models are provided an image as a visual input and a passage combined with the question as a linguistic input to predict the answer, which are referred to as IPQ baselines.

BLIP (Li *et al.*, 2022a) is a vision-language pre-training framework that effectively utilizes a synthetic caption generation along with a filter to identify noisy captions. It uses a multi-modal mixture of encoder-decoder architecture over large-scale noise-filtered image-caption data. It can transfer its learning well to both vision-language understanding and generation tasks in comparison with its predecessors. GIT (Wang *et al.*, 2022a) is a generative image-to-text transformer, which is a simple network architecture (consisting of only one image encoder and one text decoder) that achieves strong performance with data-scaling. It can be used to perform a variety of vision-language tasks including visual question answering. ViLT (Kim *et al.*, 2021) is a convolution-free pre-training approach that eliminates the need for obtaining object detection results, object tags, OCR (optical character recognition), and region features from images. ViLT is fast and parameter-efficient approach yet demonstrates strong performance in downstream vision-language tasks.

For a fair comparison of the aforementioned models, the pre-trained version of the respective model fine-tuned on the VQA dataset (Antol *et al.*, 2015a) is used to predict answers for VL-GLUE questions included in each task. Most VQA systems only take one visual and one textual input. Hence, in the case of multiple images, they are composed into a single file. Similarly, the passage and questions are concatenated to form a single language input. All models used for this experiment BLIP, GIT, and ViLT have a limit on input text tokens. To address this issue, passage inputs across all seven tasks are summarized into 50 tokens at maximum using GPT-3 (Brown *et al.*, 2020).

5.4.4 *Multi-modal Baselines (Fine-tuned)*

Finally, two fine-tuning baselines are employed- ViLT (Kim *et al.*, 2021) and VisualBERT (Li *et al.*, 2019). Firstly, their pre-trained version on VQA (Antol *et al.*, 2015a) is taken, which is fine-tuned over VL-GLUE data for each task separately. The goal here is to explore whether or not fine-tuning improves the joint-reasoning capability of models.

5.5 Results and Discussion

Table 5.4 and Figure 5.8 show the comparative performance on the seven tasks in the VL-GLUE benchmark. The models are categorized into several groups: random baseline, question-only (Q-only), passage-only (PQ-only), image-question-only (IQ-only), image-passage-question (IPQ) using multiple architectures (BLIP, ViLT, GIT), and fine-tuned models (ViLT, VisualBERT). The accuracy is used as an evaluation metric. Insights based on these results are discussed below;

Random Baseline- Lowerbound of Performance: The random baseline is reported to estimate the lower bound on performance. As different tasks in VL-GLUE have varying numbers of answer options to choose from (2 to 27), their respective probabilities to be randomly chosen and being correct can be significantly different. This can be an important factor in assessing the difficulty level of each task in VL-GLUE. From the results, it is apparent that most IPQ and fine-tuned models significantly outperform the random baseline. This indicates that the models either have learned meaningful patterns from the data or possess task-specific knowledge due to pre-training and can leverage that knowledge to answer the questions correctly. The only exception to this is Task 4, which consists of visuo-linguistic reasoning over free-form figures. This observation suggests that model pre-training is not quite helping

	Random	Q-only (GPT-3)	PQ-only (RoBERTa)	IQ-only (BLIP)	IPQ (BLIP)	IPQ (ViLT)	IPQ (GIT)	Fine-tune (ViLT)	Fine-tune (VisualBERT)
Task1	3.7%	18.9%	9.08%	30.5%	48.4%	41.1%	40.7%	44.3%	40.4%
Task2	35.4%	50.5%	51.9%	53.3%	51.24%	55.8%	57%	59.2%	51.6%
Task3	25%	22.4%	22.6%	27.3%	38.5%	33.7%	38.1%	35.1%	31.3%
Task4	31.8%	18.8%	26.3%	30.6%	35.1%	27.5%	25.6%	28.2%	25.6%
Task5	50%	61%	59.8%	60.2%	58.6%	61.1%	59.4%	61.9%	54.7%
Task6	25%	26%	23.6%	25.78%	31.89%	32.19%	31.55%	34.06%	27.88%
Task7	25%	19.1%	25.8%	45.7%	44.8%	42.6%	43.2%	45.5%	30.3%

Table 5.4: Benchmarking on VL-GLUE: Accuracy(%) for Heuristic (Random), Unimodal (GPT-3, RoBERTa, BLIP-image-only), Multimodal (BLIP, ViLT, GIT VQA), and Fine-tuned Baselines (ViLT-fine-tuned, VisualBERT-fine-tuned) for Seven Task Types

to solve instances in this task which are grounded in very specific scenarios, where there is no generic knowledge or formula sufficient to answer questions. Also, the lower performance of fine-tuned baselines for some tasks may be due to the availability of a small amount of data for training (ex. there are only 1854 instances of task 4 as per Table 5.3). The lack of existing datasets of this kind is mainly due to the diversity of images and questions that exist in this domain. As a result, it is difficult to perform synthetic data generation and requires time-consuming, cumbersome, and costly manual compilation.

Does the VL-GLUE benchmark contain bias that a model can exploit? A challenging dataset requires the model to ideally consider all the information provided as a context to arrive at an answer. To ensure that this is indeed the case, experimentation with uni-modal baselines is conducted. In this benchmark, three components formulate the context- image, passage, and question. Therefore, three uni-modal

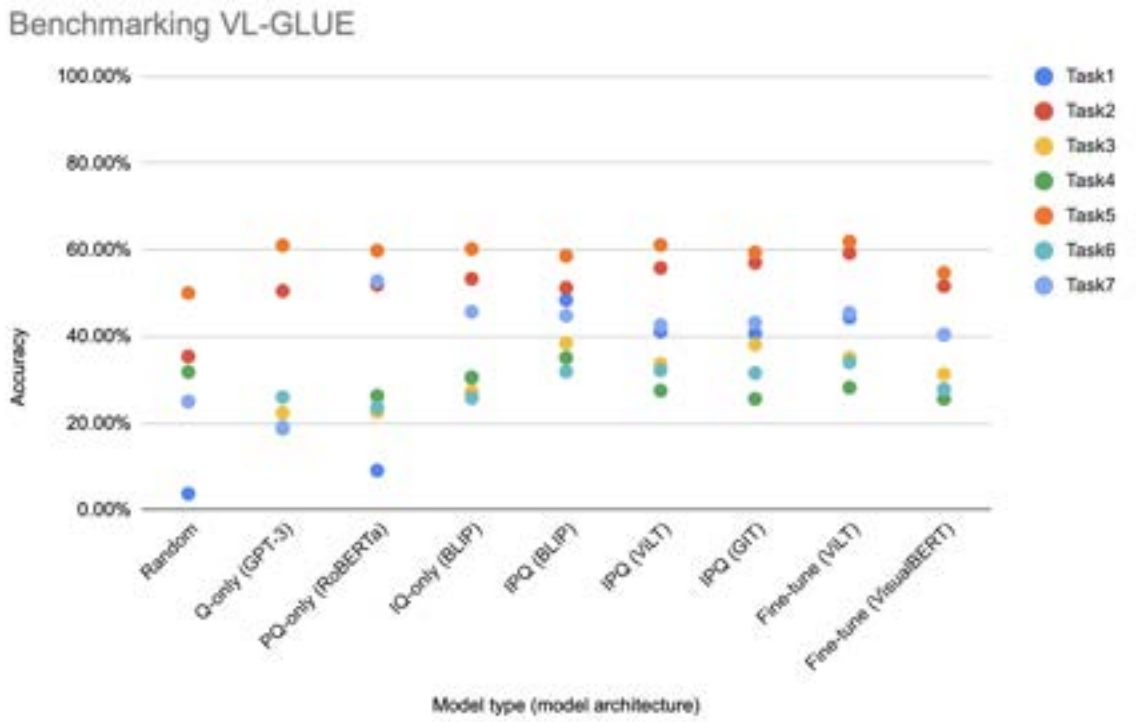


Figure 5.8: Benchmarking on VL-GLUE: Scatter Plot Representation of Table 5.4 for Visual Comparison of Accuracy Achieved by Baseline Models Employed and Variation of Model Performance Across Seven Task Types

variations are considered- (i) question-only (Q-only) which ignores both image and passage, (ii) passage and question-only (PQ-only) which ignores the image and, (iii) image and question-only (IQ-only) which ignores the passage. For most tasks, Q-only and PQ-only models demonstrate performance close to their respective random baseline performance. This indicates that for most tasks, without incorporating clues from the images it is hard to answer respective questions. The gains over random baseline for Q-only baseline are maximum for tasks 1, 2, and 5. GPT-3 (Brown *et al.*, 2020) is used to implement a Q-only baseline, which indicates that GPT-3 either has acquired some knowledge related to these tasks during pre-training and can effectively filter out distractor answer choices based on the information present in the question. This is

the most probable explanation for task 5, as this subset is based on Wikipedia, which is one of the large-scale sources that GPT-3 is pre-trained on. The achieved accuracy for the PQ-only baseline is similar to the Q-only baseline except for tasks 1, 4, and 7. The PQ-only baseline shows a 9% performance drop on task 1, whereas improves 8% and 6% over random baseline for tasks 4 and 7 respectively, which is the result of the presence of the passage modality. For tasks 1 and 7, the IQ-only baseline has notable performance improvements compared to PQ-only and Q-only models. This indicates that for both these tasks, images are a more important modality in decision-making than the passage. Overall, the relatively lower performance of the above three bias-checking baselines demonstrates that with the absence of either modality, the model performance deteriorates. This is indicative of the fact that both the benchmark and constituent tasks are challenging overall, from a visuo-linguistic reasoning viewpoint.

How good are existing multimodal models for visuo-linguistic reasoning? The IPQ models (BLIP, ViLT, GIT) demonstrate superior performance compared to random and bias-checking baselines. This highlights the importance of incorporating both image and textual information for effective answer selection. Within the IPQ category, different architectures exhibit varying performance across tasks. The BLIP model is the best among all three across all the tasks. The BLIP has the best performance on tasks 1, 4, 7, and tails by GIT with <1% accuracy difference for tasks 3 and 6. For tasks 2 and 5, GIT and ViLT are at the top of the chart respectively. Interestingly, all three models achieve relatively similar performance on tasks 5-7, whereas show the most divergence for tasks 1 & 4. We further fine-tune two models ViLT and VisualBERT on VL-GLUE data to assess whether fine-tuning pushes the performance any further. ViLT fine-tuned model is marginally better than ViLT IPQ counterpart, achieving maximum gains for tasks 1, 2, and 7. However, the limited

performance difference between these models indicates that the existing architectures do not equip models with the visuo-linguistic reasoning capability. Fine-tuned VisualBERT does not have a significant advantage over relatively newer IPQ architectures BLIP, ViLT, and GIT which benefit from both architecture advancements and access to larger training data. It is well known that the choice of architecture may have a significant influence on the model capabilities, which was the sole reason behind the adaptation of task-specific models in applications so far. However, the AI community is moving towards unified models that are general-purpose and can tackle a wide variety of tasks. We hope that VL-GLUE data will be utilized in the pre-training of next-generation vision-language models, which will equip them with visuo-linguistic capability.

Which tasks are hard to solve? Fine-tuned ViLT is the best-performing model over VL-GLUE data across all tasks. Comparing the results of this model with the random baseline, the gains achieved for tasks 1, 2, and 7 are significant (over $\sim$25%). This indicates that these tasks are relatively easy for the model to learn patterns from the respective data. One pattern among these tasks is that images are relatively simpler with a small number of objects and/or attributes present. The model's architecture along with the VL-GLUE data for these tasks are collectively helpful in achieving better visuo-linguistic reasoning capability. The fact that ViLT's performance for task 4 is worse than the random baseline performance, reiterates that this is the hardest task to solve in this benchmark. Notably, despite having the highest random selection accuracy for task 5 (50%), the gain achieved after fine-tuning is only 11%. This is an indicator that this task is challenging as well. Tasks 3 and 6 appear to be comparatively easy given the moderate performance gains. The common pattern among these two tasks is that both are built upon large-scale publicly

available internet data. It is possible that models are leveraging such information observed during pre-training, which can successfully substitute the reasoning step that bridges the given image and passage modalities.

In conclusion, visuo-linguistic (VL) reasoning is ubiquitous and an integral part of reasoning performed by humans in real-life scenarios. Motivated by this, a large-scale benchmark VL-GLUE (Sampat *et al.*, 2024a), inspired by GLUE (Wang *et al.*, 2018) and NumGLUE (Mishra *et al.*, 2022) is put together. The benchmark consists of 7 different tasks and 106k instances in total, diverse kinds of images (from simple rendered images to charts, freeform diagrams, and images requiring identification of named entities) and spans across multiple domains (education, politics, sports, history, cooking, and day-to-day activities). The experimental results demonstrate that VL-GLUE is a challenging benchmark for the latest large-scale vision-language models, obtaining poor scores not only in zero-shot settings but also after fine-tuning. This effort towards collection/expansion and standardization of existing VL datasets is aimed to serve as a resource for further research in this area and the development of AI models with superior multi-modal reasoning capabilities. The code and links to the data are available at `https://github.com/shailaja183/VL-GLUE.git`.

Chapter 6

A ZERO-SHOT APPROACH TO LEARN IMAGE-SPECIFIC COMMONSENSE CONCEPTS ABOUT ACTIONS

6.1 Motivation

From a very early age, humans can form a mental representation of surrounding objects through a combination of active learning (first-hand experience and imitation) and passive learning (reading from books, watching videos, conversing with other people, etc.). As they grow older, they are able to make more complex and abstract inferences, including the ability to predict intentions, anticipate future states of the world, and respond to imaginary situations (Moore, 2013). Over time and with experience, humans build a cumulative knowledge repository of reusable concepts that enable their interaction with the world seamlessly.

Moreover, people's interaction with the world is heavily goal-driven. Humans capitalize on their knowledge about objects (such as their properties, relationship with other objects, and affordances) and perform necessary actions over them to accomplish desired goals. That being said, a significant amount of commonsense knowledge that humans use in their day-to-day lives revolves around actions. For example, one can easily perceive when a stack of dishes or a Jenga tower will topple; one can predict whether the object is firmly attached to the surface or free to be lifted; and one can estimate to what extent the grocery bag should be filled so that it does not tear or crush the contents (Battaglia *et al.*, 2013).

Commonsense reasoning about actions is crucial for humans to navigate in the world, as highlighted in several works- (i) to visualize what sequence of actions will

lead to the goal and to reduce exploration (Murugesan *et al.*, 2021); (ii) to perform look-ahead planning and determine how current actions will affect future states of the world (Juba, 2016); (iii) to explain observations in terms of what actions may have taken place (Baral, 2010); and (iv) to diagnose faults i.e. identifying actions that may have caused undesirable situations (Baral, 2010).

As autonomous agents are being developed that would assist humans in performing everyday tasks, they would also have to interact with complex environments and possess the commonsense to resolve situations that may arise. The importance of this research problem has been identified since the early days of AI (McCarthy *et al.*, 1960; McCarthy, 1963; McCarthy and Hayes, 1969). Since then, there has been a lot of progress and active efforts in curating large-scale commonsense resources and in building systems that can tackle varied aspects of action reasoning. However, state-of-the-art models (including large language models) face numerous challenges with tasks that require reasoning, which is quite trivial for humans (Li *et al.*, 2022b; Valmeekam *et al.*, 2022; Bian *et al.*, 2023). One key reason for this performance gap between humans and AI models is the coarse-grained nature of most existing commonsense benchmarks e.g., a person cooks if they are hungry, a person calls the police in order to be lawful, a person will smile if someone complements them (Sap *et al.*, 2019).

Commonsense concepts related to actions are much more complex behind the scenes than obvious surface-level correlations, which in turn require consolidated physical, spatial, temporal, visual, and causal understanding. Such concepts include pre-conditions (e.g. necessary ingredients and kitchen equipment), executability (e.g. liquid objects can be poured, solid objects can be cut), effects (e.g. straining will remove excess water), high-level goals (e.g. to make an omelet) and temporal dependency (e.g. crack eggs before beating or whisking). This work (Sampat *et al.*, 2024c) aims to address this challenge by contributing the following:

- A novel multi-modal task to learn concepts about actions (pre-conditions, effects, goals, and past/future actions) from images is proposed.
- A zero-shot approach to discern knowledge about such action-related concepts from the large language model GPT-3 is developed.
- A dataset of 8.5k images and 59.3k inferences about actions grounded in images is collected from an annotated cooking-video dataset, which is used to evaluate the quality of language model-generated inferences.
- Ablations are performed using diverse input prompts and various methods to incorporate visual information are explored. Quantitative/qualitative results of the experiments are reported and compared with the performance of existing VQA systems.

6.2 ActionCOMET Task Overview

There exists an array of well-defined tasks and datasets in the vision-language domain (Yeo *et al.*, 2018; Storks *et al.*, 2021; Bisk *et al.*, 2020b; Dalvi *et al.*, 2019; Yang *et al.*, 2021b; Gao *et al.*, 2018; Isola *et al.*, 2015) that focus on reasoning about actions (Sampat *et al.*, 2022b). However, most of the aforementioned benchmarks focus only on a particular aspect of action understanding among physical, spatial, temporal, visual, or causal understanding. By limiting the focus to a particular aspect, the model evaluation process remains simple and transparent. However, models trained on a particular aspect lack the big picture and fail to understand complex interrelations with other aspects that were not observed during training. In other words, a model trained over causality of actions-related data lacks an understanding of how an object should be kept spatially to perform an action or how the world will

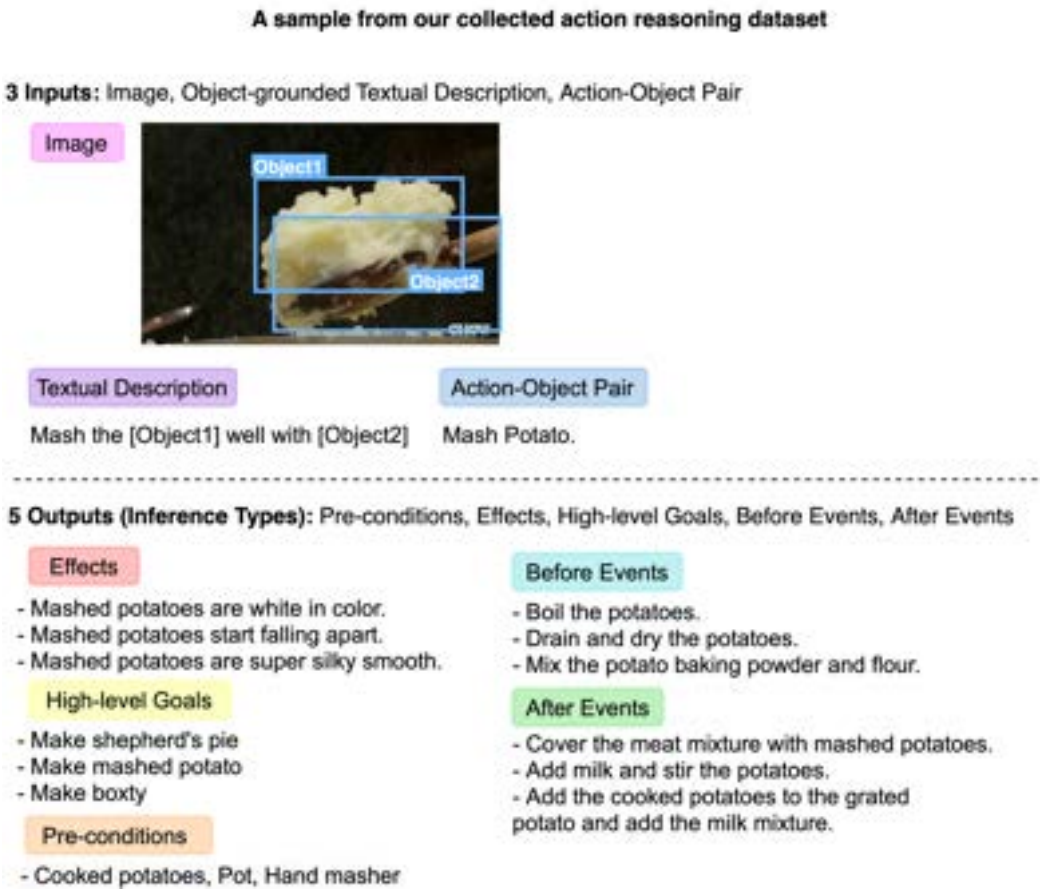


Figure 6.1: Fine-grained Action-centric Commonsense Generation Addressed in This Work: (Top-left) Input to the Model: an Image, a Textual Description of an Image, and an Action-object Pair of Interest (Top-right) Five Types of Action-related Inference That Needs to be Predicted Using a Vision-Language Model: Effects, High-level Goals, Pre-conditions, Before Events and After Events (Bottom) the Overview of the Proposed ActionCOMET Model Demonstrating How Different Inputs are Processed and the Inference is Generated

change visually after the action is complete, and so on. Motivated by the deficiency of benchmarks in the existing literature that supports consolidated physical, temporal, visual, and causal understanding, in this work, a new task is proposed. In particular, the task is to narrate a dynamic story of an image from action-related aspects, as per the example shown in Figure 6.1. The dynamic story consists of seven major commonsense inference components:

1. **Textual Description:** This is equivalent to a dense image caption narrating objects of interest and actions being performed in the image. It would be of the form 'cracking [Object1] using [Object2]' if the egg and fork are detected in the image respectively. By leveraging actions and objects detected in the images into this description, the goal is to generate inferences that are conditioned on the content of images.

2. **Action-Object Pair:** This component contains a lemmatized primary verb and noun pair for the current event (e.g. crack egg). This can be quite similar to textual description but extracted from YouCook2 annotations instead of the visual input. This component ignores secondary verbs and other supporting objects in the event that do not undergo the action directly. For example, if the image consists of an action where a fork is used to crack the egg. In this case, the fork does not undergo cracking action itself therefore 'crack fork' is not considered as a valid action-object pair.

3. **High-level Goal:** This component describes the motivations or intentions of a person/agent to perform a particular action over an object (e.g. action knead the dough is associated with the goal of baking the bread). The inclusion of this inference type is inspired by goal-driven human psychology to decide the course of action.

4. **Pre-conditions:** The set of actions you can perform on an object depends on the existence of other objects in the surrounding environment. For example, the presence of a knife or a scissor in order to perform a cut action. This component plays an important role in equipping models with the capability to infer commonly used objects and ingredients that play a role in performing actions.

5. **Effects:** This inference describes the states of the object after an action is successfully performed over it. The state of the object can be described in the form of visual appearance (e.g. brown, puree, melted) or in terms of attributes it possesses (e.g. raw, soft, crisp). For example, peeling orange action removes the outer skin of an orange, or frying a potato leads to the potato being golden or crisp.

6. **Before Event:** To reason about the dynamic situation of the image in the past, this inference type is used. This includes either the actions that might have taken place leading to the condition visible in the given image (e.g. peeling potato might have taken place before slicing potato if the skin is not visible in the image) or actions that are commonly performed before reaching the state shown in the image (e.g. oven is pre-heated or water is boiled in a pot).

7. **After Event:** To reason about the dynamic situation of the image in the future, this inference type is used. This includes either the actions that are likely to take place after the condition visible in the given image for a specific goal (e.g. frying potato may take place after slicing if the goal was to make french fries), or events involving other objects (e.g. season the potato slices or coat them with corn flour).

The commonsense reasoning task is plausible in most cases i.e. one can come

up with multiple reasonable commonsense conclusions depending on the information provided, assumptions they make, and past knowledge (Davis and Marcus, 2015). For example, the action cut a cauliflower can be associated with many possible high-level goals such as- to make a cauliflower steak, soup, curry, pizza crust, and many more. Humans might prefer some inferences over others depending on their personal preference or their knowledge of dishes that use cauliflower as an ingredient. Similar reasoning can be applied to other inference types. To account for plausibility, each of the above commonsense components is framed as a set of one or more inferences.

6.3 ActionCOMET Dataset Creation

6.3.1 Data Collection

As discussed above, given an image, the task is to generate inferences of five kinds- (i) pre-conditions, (ii) high-level goals, (iii) effects, (iv) before events, and (v) after events. To support the development and evaluation of models that demonstrate consolidated physical, temporal, visual, and causal understanding by generating the above five kinds of inferences, a dataset is compiled based on the cooking domain. In cooking demonstrations, people typically narrate the recipe preparation process step-by-step including the details about objects that are necessary to perform an action, how to perform the action, how an object should look after the action is performed, and in what order actions should be performed. Therefore, the cooking domain is preferred to collect data for the proposed task, which is rich in terms of diverse activities and associated commonsense inferences. Such a dataset would strongly reflect how humans operate in this world and achieve desired goals by performing actions.

An existing annotated video dataset and corresponding video transcripts are used

as a resource to collect data for this task. A pipeline based on multiple off-the-shelf NLP tools (including co-reference resolution, dependency parser, and reading comprehension models) is set up to extract data for all five inference types, with minimal human intervention. This results in a dataset that consists of plausible inferences about various actions that are viable from a human's perspective. The following section describes the data collection process in detail.

The existing YouCook2 dataset (Zhou *et al.*, 2018b) is leveraged for this purpose, which is a collection of large task-oriented instructions (about the cooking domain) that is developed to improve video understanding tasks. Their train+val partition is used, which contains annotations about 1790 videos available on YouTube covering 89 unique recipes. Each video in the dataset is divided into multiple procedural steps or segments $\{S_1, S_2,..., S_N\}$, which represent key events occurring in the video. For each segment S_i, i $\in \{1,2,..,N\}$, the dataset provides annotated temporal boundaries (i.e. B_i^{start} and B_i^{end}) and a descriptive sentence T_i, as per example shown in Figure 6.2. The dataset also provides bounding box annotations $O_i = \{o_i^1, o_i^2, ..., o_i^M\}$ where o_i^1, o_i^2, ..., o_i^M are objects present in segment S_i (Zhou *et al.*, 2018a). There is a RecipeID assigned to each video, denoting the type of the recipe (among 89).

The YouCook2 dataset (Zhou *et al.*, 2018b) does not provide actual videos in their dataset. So using the yt-dlp[1] python package, video-only data is obtained, and using youtube-transcript-api[2], the corresponding audio transcript is extracted. Afterward, the obtained video is segmented into clips $\{C_1, C_2,..., C_N\}$ and audio into $\{A_1, A_2,..., A_N\}$ using the segment duration (B^{start}, B^{end}) so that it can be used in sync with rest of the annotations. During this step, 189 videos were not accessible as they were no longer available on YouTube. Similarly, 395 transcripts could not be obtained

[1] `https://github.com/yt-dlp/yt-dlp`

[2] `https://pypi.org/project/youtube-transcript-api/`

YouTubeId: 4eWzsx1vAi8
RecipeId: 101 (BLT Sandwich)

Segment (S)	Start/End (B^{start}, B^{end})	Textual Description (T)
S1	(00:21, 00:51)	Grill the tomatoes in a pan and then put them on a plate.
S2	(00:54, 01:03)	Add oil to a pan and spread it well so as to fry the bacon.
S3	(01:06, 01:54)	Cook bacon until crispy, then drain on paper towel.
S4	(01:56, 02:40)	Add a bit of Worcestershire sauce to mayonnaise and spread it over the bread.
S5	(02:41, 03:00)	Place a piece of lettuce as the first layer, place the tomatoes over it.
S6	(03:08, 03:15)	Sprinkle salt and pepper to taste.
S7	(03:16, 03:25)	Place the bacon at the top.
S8	(03:25, 03:28)	Place a piece of the bread at the top.

Figure 6.2: Sample 4eWzsx1vAi8 From Train+Val Partition of YouCook2 Dataset with Annotated Procedure Steps

due to the video being unavailable or closed captioning turned off by the user. The remaining 1601 videos are processed through the pipeline shown in Figure 6.3.

As a first step, co-references present in the textual descriptions T_i are resolved. For the BLT Sandwich example (Figure 6.2), segments S_1, S_2 and S_4 and S_5 contain co-references (them/it) which refer to tomatoes, oil, mayonnaise, and lettuce respectively. In particular, F-coref's (Otmazgin *et al.*, 2022) Spacy component is used for this purpose. Further, textual descriptions (T_i with co-references resolved) are passed through the dependency parser to extract all (Verb, Ingredient) pairs that directly connect in a dependency tree. In the above example, from S_1 two such pairs can be obtained- (Grill, Tomato) and (Put, Tomato). This process is repeated for all the segments in a video. Once all such pairs are collected for the entire dataset, only the pairs that have at least 10 occurrences of verbs or ingredients (independent of

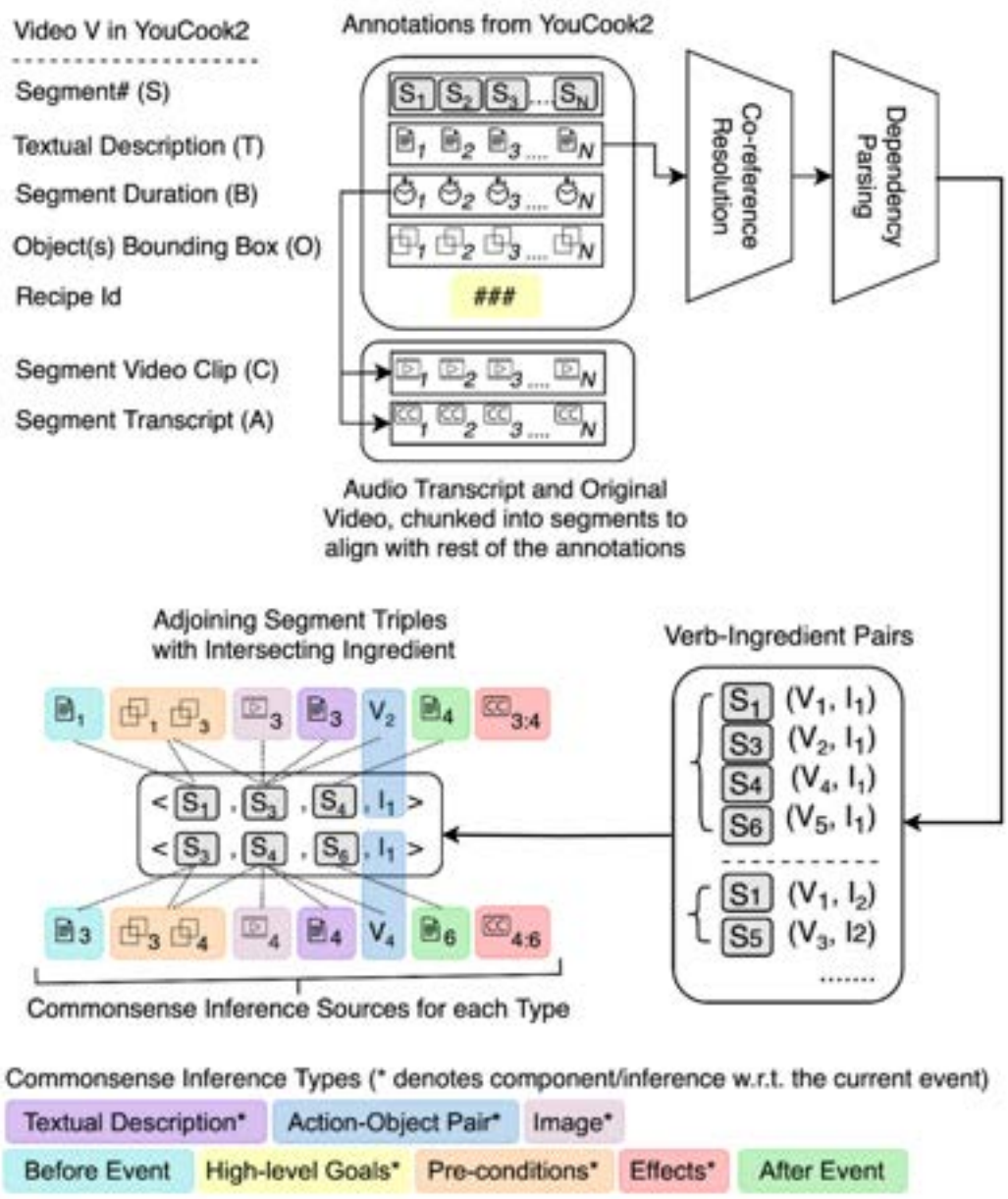


Figure 6.3: The Data Preparation Pipeline That Leverages YouCook2 (Zhou *et al.*, 2018b) Annotations to Extract Commonsense Inference About Actions (Best Viewed in Color)

each other) are retained. Such filtering is employed to reliably collect commonsense inferences around verbs/ingredients and avoid any noise or rare utterances. Spacy's transformer-based en_core_web_trf pipeline (Honnibal *et al.*, 2020) is used, which is state-of-the-art for dependency parsing.

After obtaining all verb-ingredient pairs based on the above criteria, they are grouped by ingredients and sorted in ascending order by segment number to make sure their temporal ordering is preserved. For each ingredient, adjoining segment triplets <S_i, S_j, S_k> are collected such that i<j<k. Here, S_j would be considered as a current event, S_i and S_k being past and future events respectively. For the BLT sandwich example, the ingredient 'bacon' appears in segments S_1, S_2, and S_7 hence a segment triplet <S_1, S_2, S_7> can be created. This denotes there are three actions fry $\rightarrow$ cook $\rightarrow$ place is performed over the bacon. If 'cook bacon' is considered a current event, 'fry bacon' and 'place bacon' would make the past and future events respectively.

It is also possible that the ingredient may undergo processing in more than three segments, for example, the case in Figure 6.3. There are four segments S_1, S_3, S_4, S_5 with ingredient I_1. Here, two adjoining segment triplets can be created i.e. <S_1, S_3, S_4> and <S_3, S_4, S_6>. More generally, if an ingredient occurs in K different segments, where K<N, then K-2 segment triplets will be obtained. For each triplet obtained, annotations {T, O, C, A, RecipeID} are paired with an image corresponding to the current event. Finally, the commonsense inferences described in Section 6.2 are collected as follows;

1. **Textual Description:** T_j with object grounding corresponding to segment S_j would be considered as a ground-truth textual description of the current event. For testing instances, existing image captioning systems will be used to obtain textual descriptions.

2. **Action-Object Pair:** Verb V and Ingredient I associated with segment S_j would be considered as an action-object pair for the current event.

3. **High-level Goal:** The recipe name lookup is performed using the RecipeID from the annotation which is substituted into a text template 'Make/Cook/Prepare ¡recipe name¿' to generate this inference.

4. **Image:** The middle frame of the video clip C_j corresponding to segment S_j is captured, which is a visual representation of the action-object pair for the current event.

5. **Pre-conditions:** The bounding box annotations O_i and O_j corresponding to past event and current event are used together to form pre-condition inference. For example, consider the past event 'potato is peeled using a peeler' and the current event 'potato is cut into pieces using a knife over a chopping board'. In this case, {potato, peeler, knife, chopping board} form pre-conditions for the current event.

6. **Effects:** Audio transcripts between the current event and future event i.e. $A_{i:j}$ is used to extract effects. Typically, in cooking videos, the person demonstrating the recipe will act first (i.e. current event) and then (but before the next event) describe how the object should look after the action, in terms of color, shape, or other attributes. Specifically, the audio transcript $A_{i:j}$ is fed into a reading comprehension model, and templated questions are asked such as 'What color/texture/shape is the <ingredient>?' and 'What attribute/property is related to <ingredient>?'. The RoBERTa-base model (Liu *et al.*, 2019), fine-tuned over SQuAD2.0 dataset (including unanswerable questions) (Rajpurkar *et al.*, 2018) is used.

7. **Before Event:** T_i corresponding to segment S_i (pre-cursor to the current event S_j) is used as a textual description of the before event.

8. **After Event:** T_k corresponding to segment S_k (following event to the current event S_j) is used as a textual description of the after event.

As a final step, inferences obtained from multiple videos that have common action-object pairs are combined to support the plausible nature of commonsense. For example, 'mash potato' was a part of three different recipes (shepherd's pie, mashed potato, and boxty), hence their inference across each type are combined as shown in Figure 6.1 (bottom).

Note that, in this work, the underlying interest is to explore how well state-of-the-art language models and vision-language models can reason about actions. The proposed model is zero-shot in nature, therefore there is no explicit training or fine-tuning done with respect to the dataset collected using the above process. By comparing the inferences stored in this dataset with model-predicted inference, the goal is to distinguish aspects that existing models do well on or struggle with. However, this resource is fairly large (with an average of 11k inferences grounded in 8.5k images per inference type) and can be leveraged for training specific models that can do better reasoning about actions.

6.3.2 Dataset Statistics

Table 6.1 summarizes the statistics about the data obtained using the process outlined in Section 6.3. The size of the resulting dataset is relatively smaller in comparison with existing benchmarks such as VisualCOMET (Park *et al.*, 2020),

Annotated Input Type	**#Samples**
Videos	1601
Images*	8522
Textual Descriptions*	8522
Recipe Types	89
Actions-Objects* ($\geq$10 occurrences)	
Unique Objects	176
Unique Actions	93
Commonsense Inferences	
High-level Goals*	10341
Pre-conditions*	17209
Effects*	6428
Before Events	12665
After Events	12665

Table 6.1: Statistics for Data Collected in This Work (* Denotes Component or Inference Desired With Respect to the Current Event)

which was completely crowd-sourced by human workers[3]. The proposed pipeline is easily reusable and scalable, even for domains beyond cooking. More importantly, the data collection process in this work is fully automated and does not require human annotation.

[3]$4/image paid to AMT workers * $60k images = $240k was the total cost of data collection for VisualCOMET as reported in Park *et al.* (2020)

6.4 Experiments

To evaluate the action reasoning task proposed in this work, an experiment with two sets of models is conducted; generative VQA models (including a few which are trained on datasets that require external knowledge to answer questions) and a large-language model GPT-3. Both sets of models used here are zero-shot in nature i.e. no explicit training or fine-tuning is performed. The dataset obtained (using the method described in Section 6.3) is utilized as ground truth to compare model predictions using reference-based automatic evaluation measures. Refer to Section 6.5 for details about the evaluation process. By analyzing the results from the above experiments, insights are sought on how well large-scale pre-training approaches are equipped with knowledge about fine-grained aspects of action understanding.

6.4.1 Can Existing VQA Models Reason About Actions?

In this section, experiments are conducted with respect to existing visual question answering (VQA) models to gauge their ability to reason about various aspects of actions included in the proposed task (described in Section 6.2). In particular, experiments with three variants of BLIP (Li *et al.*, 2022a) model are done that include pre-training over VQAv2 (Goyal *et al.*, 2017), OKVQA (Marino *et al.*, 2019), A-OKVQA (Schwenk *et al.*, 2022). Note that a significant portion of VQAv2 can be answered based on immediate visual content, whereas OKVQA and A-OKVQA benchmarks require a broad base of commonsense and world knowledge to answer questions about the images. These models are used in generative question answering setting given an image, using the implementation by Li *et al.* (2023).

For each model variant described above, the following set of templated questions are posed:

- Q1: Which action is being performed? → A1
- Q2: Which object is being <A1>? → A2
- Q3: Which other object is participating in <A1> action?
- Q4: Which action will happen after the <A1> <A2>?
- Q5: Which action was performed before the <A1> <A2>?
- Q6: Which recipe can one make by <A1> <A2>?
- Q7: How will the <A1> <A2> look like?

Specifically, the model is asked to predict the action that is about to happen, recently happened, or is in progress in the image (in Q1). The answer to Q1 is denoted by A1. Then, in Q2, the model is asked to identify an object on which the action A1 is being performed and store the answer as A2. To obtain pre-requisites, Q3 is asked i.e. which other objects are participating in the action. For most likely past and future events, Q4 and Q5 are posed respectively by asking what action was performed before and what action will happen after the current event. Then the name of the recipe is asked which would require performing action A1 on object A2, which would be analogous to the high-level goals. Finally, by asking the model to describe object A2 after performing action A1, inference for effects will be obtained.

For example, consider a scenario where for a particular image, the predicted answer A1 for Q1 is 'cut' and the predicted answer A2 for Q2 is 'potato'. In that case, the subsequent questions will be 'Which other object is participating in cut action?', 'Which action will happen after the cut potato?', 'Which action was performed before the cut potato?', 'Which recipe can one make by cut potato?' and 'How will the cut

potato look like?'. The predictions of VQA models posed with the above set of questions are evaluated against inferences collected in this dataset using text generation metrics.

6.4.2 Can GPT-3 Reason About Actions?

A human or a model's understanding of actions can be characterized through five tightly coupled aspects, which include pre-conditions, effects, high-level goals, and a set of actions that are likely to precede or follow. The recent developments in large language models (LLMs) indicate that such models (Brown *et al.*, 2020; Touvron *et al.*, 2023; Penedo *et al.*, 2023) possess a vast amount of information about the world due to training on massive corpora (Petroni *et al.*, 2019; Roberts *et al.*, 2020; Hendrycks *et al.*, 2020). In this section, the goal is to explore the capabilities of GPT-3 (Brown *et al.*, 2020) to reason about actions, which can be broken down into four key questions:

1. How to formulate an input prompt for GPT-3 that can generate inferences related to actions shown in an image?

2. Does auxiliary information about visual inputs (e.g. captions) improve the generation?

3. Do variations of textual instructions used to distinguish the five inference types affect the generation?

4. How well are the GPT-3 generations in comparison with annotations collected from the cooking videos described in Section 6.3?

To this end, a workflow is developed considering the above research questions.

Formulating an Input Prompt for GPT-3 For each image in the proposed dataset, five inferences are to be generated corresponding to (i) pre-conditions, (ii) high-level goals, (iii) effects, (iv) before events, and (v) after events. The natural first step is to extract a sequence of visual embeddings V representing the image and identify object regions in the image. The Region of Interest (RoI) Align features (He *et al.*, 2017) from Faster-RCNN (Ren *et al.*, 2015b) are used as visual embedding and passed through a non-linear layer to obtain the final representation for an image or each detected object. The final sequence of representations is referred to as V = $\{v_0, v_1, ..v_m\}$ where m is the number of objects detected.

In many vision-language tasks, the presence of auxiliary information (provided as a direct input during prediction or through supervised training) has shown improvements in the downstream task performance. Therefore, two auxiliary inputs are considered for this task: textual description of the image (TextDesc) and ground-truth action-object annotation (AO Pair).

The task of VisualCOMET (Park *et al.*, 2020) about generating commonsense inferences (of events and intents) from people-centric images is quite similar to the focus of this work. In VisualCOMET, a 'Person Grounding' technique was introduced, where each person detected in the image is assigned special tags [Person1], [Person2], and so on. These special tags are referred to within the event descriptions (e.g. '[Person1] is pointing towards [Person2]') to have tighter integration between the images and textual components. Park *et al.* (2020) demonstrated that when this technique was used, their inference generations were improved. This technique is adapted in this work and referred to as 'Object Grounding (OG)' as images in this work involve numerous objects that either undergo different actions or help in the execution of actions. In this case, object tags are used such as [Object1], [Object2] detected in the image, which are referred to in the text descriptions (TextDesc) ex.

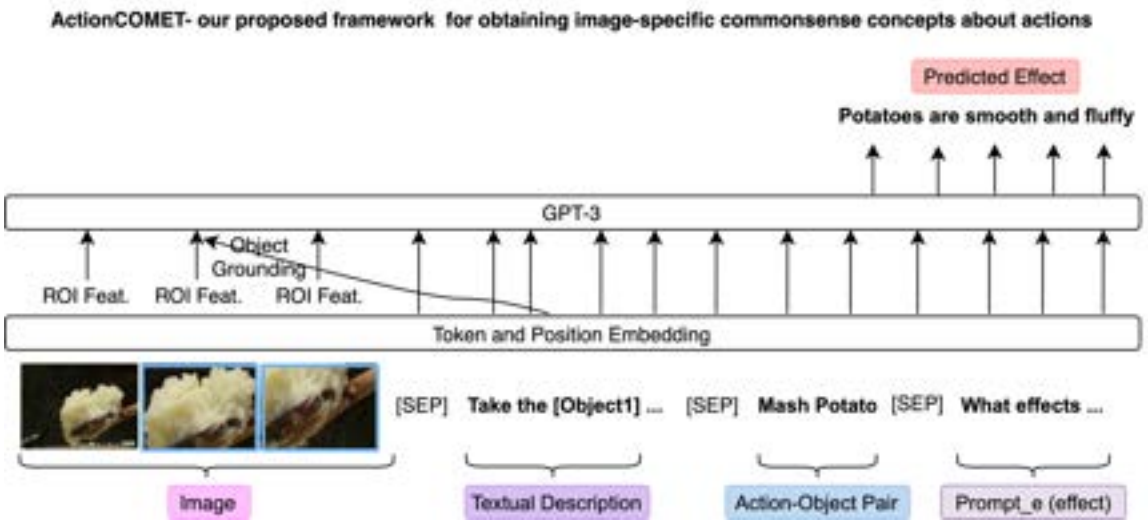


Figure 6.4: Overview of the Proposed GPT-3 Based Pipeline, Which is a Zero-shot Approach for Generating Fine-grained Commonsense About Actions Conditioned on the Images

'chop [Object1] using [Object2]'.

To generate five kinds of inferences related to actions, five text prompts are formulated: Prompt_p (pre-conditions), Prompt_e (effects), Prompt_g (goals), Prompt_b (before events), and Prompt_a (after events). For example, in order to obtain the effects of the action shown in a given image, the text prompt 'Some results of performing this action include' can be used. Similarly, to query GPT-3 about the pre-conditions, the text prompt 'List down things without which one can not perform this action' can be used.

The four inputs described above are combined in the following order; the image features along with object detection (Image), a textual description of the image (TextDesc), ground-truth action-object annotation (AO Pair), and text prompt corresponding to inference i (Prompt_i), which are separated by [SEP]. The pipeline that uses GPT-3 and takes the above four inputs to generate inference related to actions

can be visualized in Figure 6.4.

Ablations Involving Vision-related Components In many vision-language tasks, the presence of auxiliary information (provided as a direct input during prediction or through supervised training) has shown improvements in the downstream task performance. For example, works of Wu *et al.* (2019); Huang *et al.* (2021); Lin and Parikh (2016); Luo *et al.* (2021) leveraged image captions, object labels, object attributes, or external knowledge as an additional input beyond image features to improve the visual question answering task. In experiments, the fixed model architecture is used as described in the above. The ablations are performed on different combinations of the inputs available, which include an action-object pair (AO Pair), a textual description of the current event (TextDesc), and an Image. The impact of object grounding (OG) is also assessed in cases where Image input is incorporated. As a result, 10 different combinations can be formulated (as listed in Table 6.3). For each modality combination, the generated inferences are evaluated on the collected data using BLEU-2 (B), METEOR (M), CIDER (C), Acc@50, Uniqueness, and Novelty metrics. The effect of variations in visual inputs is reported in Table 6.3.

Ablations Involving Textual Components Several research works (Brown *et al.*, 2020; Jiang *et al.*, 2020; Chen *et al.*, 2022; Wang *et al.*, 2022b) have shown that prompting GPT-3 with different variations of the instructions may have an impact on the generation process. As shown in the bottom part of Figure 6.1, a natural language prompt is used to instruct the model to specify which type of inference (among pre-condition, goal, effect, before action, or after action) is to be predicted. In this section, the effect of using different prompts on the downstream performance is evaluated. In particular, four variations of each inference type are considered- As-

sertive (sentence completion), Imperative, Interrogative, and Assertive (structured response). These prompt types were inspired by fundamental sentence constructions and language tones that are often used in human languages i.e. English. Existing works Yin *et al.* (2024) have shown that LLMs are sensitive to such sentence variations and tones. The prompt variations incorporated in this study are summarized in Table 6.2.

Assertive (sentence completion) prompt refers to instruction that provides the model with some context about the type of desired inference (goal/pre-condition/effect etc.) and expects the model to complete the remaining sentence with appropriate inference. Pp1/Pe1/Pg1/Pb1/Pa1 prompts (referred in Table 6.2) are of assertive (sentence completion) kind. The imperative prompt refers to the sentence formulation which proposes a model to return a particular response in an instructive or requesting tone. It typically starts with 'Describe..' and covers prompts with suffix 2 (e.g. Pp2) in Table 6.2. The interrogative prompts, as the name suggests, are formulated as Wh-questions beginning with 'What are some ...'. Finally, the assertive (structured response) prompt instructs the model to return an itemized or enumerated answer instead of the free-form natural language response by explicitly stating the model to 'List down' its response.

Inference Generation The pre-trained GPT-3 (Brown *et al.*, 2020) is used for natural language generation, conditioned on the four inputs. Particularly, the text-davinci-003 variant is used in text completion mode with a maximum sequence length of 32. For decoding, nucleus sampling (Holtzman *et al.*, 2019) with p = 0.9 allows diverse text generation contrary to beam search, following Park *et al.* (2020). Visual features for image and object embeddings use ResNet101 (He *et al.*, 2016) backbone pre-trained on LVIS (Gupta *et al.*, 2019). The maximum number of visual features

Inference	Variations of Prompts Considerded
Prompt_p (pre-condition)	Pp1: Some concepts that are required to perform this action are Pp2: Describe a list of necessary conditions to execute this action Pp3: What are some pre-requisites related to this action? Pp4: List down things without which one can't perform this action
Prompt_e (effect)	Pe1: Some results of performing this action include Pe2: Describe changes caused by performing this action Pe3: What effects will be produced as a result of this action? Pe4: List down the consequences if one performs this action
Prompt_g (goal)	Pg1: Some objectives related to this action include Pg2: Describe intents of people that are performing this action Pg3: What are some high-level goals associated with this action? Pg4: List down recipes one can prepare by performing this action
Prompt_b (before action)	Pb1: Some actions that person must have performed before this action are Pb2: Describe which actions might have taken place in past Pb3: What are some actions that take place before this action? Pb4: List down some actions that preceded this action
Prompt_a (after action)	Pa1: Some actions that person will perform after this action are Pa2: Describe which actions are likely to take place in future Pa3: What are some actions that take place after this action? Pa4: List down some actions that will follow this action

Table 6.2: GPT-3 Prompt Variations Considered for Five Action Inference Types (Pre-conditions, Effects, Goals, Preceding, Following Actions) in the Proposed Dataset

is set to 15.

6.5 Results and Discussion

Evaluation Metric The automatic evaluation measures that are used to evaluate the quality of inference sentences are discussed in this section. Particularly, BLEU-2 (Papineni *et al.*, 2002), METEOR (Denkowski and Lavie, 2014), and CIDER (Vedantam *et al.*, 2015) automatic metrics commonly used for image captioning tasks are used to evaluate the generated inferences. Inspired by the perplexity metric used for ranking inferences in visual dialog tasks, this metric is incorporated as well. Negative/distraction inferences are appended in such a way that there are 50 candidates to choose from. The candidates are ranked using the perplexity score, and average accuracy of the retrieved ground truth inferences (A@50) is measured. Note that perplexity may not be a perfect measure to rank the sentences, but good language models should still be able to filter out inferences that do not match the content in image and event at present. Lastly, the diversity of sentences is measured, in order to reward the model for being creative and discourage models from making same predictions over and over. The ratio of unique sentences and total number of sentences generated is considered as Unique metric. The ratio of generated sentences that are not in the training data and the total number of sentences constitutes the Novelty metric.

Automated Evaluation Table 6.3 shows the experimental results of testing with different combinations of input modalities. The following observations can be inferred: First, model that incorporates both visual and textual (Image + TextDesc + AO pair + OG) modalities outperform models trained with only one of the modalities (TextDesc + AO pair; Image + OG) in every metric, including retrieval accuracy and

Modalities related to visual inputs	B	M	C	A@50	Unique	Novel
Image	7.34	8.12	6.55	13.60	23.14	45.30
Image + OG	9.02	10.45	8.67	16.37	26.30	48.26
AO Pair	6.77	9.01	7.66	11.35	32.36	56.12
TextDesc	7.43	8.74	6.80	13.52	34.00	53.18
AO Pair + TextDesc	7.49	8.68	6.91	13.47	35.23	52.84
Image + TextDesc	10.32	11.18	9.85	18.55	37.12	54.11
Image + AO Pair	15.45	17.10	16.85	21.28	35.33	50.45
Image + TextDesc + AO Pair	15.76	17.08	16.77	22.01	34.61	51.06
Image + TextDesc + OG	**17.44**	**18.56**	**17.20**	**24.08**	**36.31**	**49.56**
Image + TextDesc + AO Pair + OG	**17.21**	**18.67**	**17.35**	**23.57**	**37.31**	**51.56**

Table 6.3: Ablations of Baseline Model from Visual Modalities Perspective: Reported Results are for the Validation Set Using Metrics BLEU-2 (B), METEOR (M), CIDER (C), Acc@50, Uniqueness and Novelty of Generated Inference (TextDesc is Equivalent to the Caption of the Image, AO Pair Denotes Ground-truth Action-object Pair Shown in the Image and OG Denotes the Use of Object Grounding)

diversity scores. This indicates that the task requires visual information is crucial and by incorporating textual descriptions relavant to the image can help to achieve higher quality inference. This reinforces the fact that the task at hand is of multi-modal nature.

Second, incorporation of action-object pair (AO pairs) seems to be more effective in comparison with textual description of the current event (TextDesc) when combined with images from the results perspective. As described earlier, AO pairs only incorporate one object at a time, which enables model to remain focused while generating inference about a given action-object combination. However, there might be another objects and actions that may passively participate in the action but

Prompt Variation	B	M	C	A@50	Unique	Novel
Pp1	17.09	18.43	17.65	23.31	37.48	51.78
Pp2	**18.33**	**20.41**	**19.19**	**24.28**	**36.78**	**50.55**
Pp3	18.19	18.67	18.48	23.01	39.22	51.99
Pp4	17.57	19.42	19.79	24.04	37.44	50.35
Pe1	13.36	16.22	15.44	17.81	36.18	42.55
Pe2	12.69	15.57	14.81	17.55	37.12	40.24
Pe3	13.25	16.07	15.56	18.11	36.26	43.78
Pe4	**15.34**	**17.34**	**16.66**	**18.27**	**35.42**	**43.35**
Pg1	15.27	17.45	16.89	20.01	38.58	40.66
Pg2	15.56	18.35	17.73	18.87	39.47	41.23
Pg3	14.80	17.54	15.88	19.41	40.20	42.14
Pg4	**16.23**	**18.89**	**17.77**	**20.24**	**37.20**	**43.67**
Pb1	21.05	23.44	20.81	26.67	30.34	45.13
Pb2	20.66	22.68	19.72	24.12	31.59	41.47
Pb3	**21.34**	**24.05**	**21.67**	**25.78**	**30.56**	**43.25**
Pb4	19.08	21.82	22.01	25.12	31.22	42.42
Pa1	17.87	20.11	17.89	24.31	33.01	46.22
Pa2	16.46	19.87	17.88	23.16	30.29	42.45
Pa3	**19.03**	**22.41**	**20.21**	**25.37**	**34.11**	**45.27**
Pa4	17.56	21.61	19.33	24.03	31.29	43.65

Table 6.4: Ablations of Baseline Model from Textual Prompts Perspective

do not undergo any effects themselves. Such broader context can only be understood through longer text modalities such as TextDesc. There is only a marginal performance improvement when both textual modalities are present together i.e. Image+AO Pair+TextDesc combination. Finally, when working with longer texts (TextDesc), the object grounding (OG) trick gives a performance boost to the model across all metrics.

Table 6.4 shows the experimental results of testing with prompt variations. For this experiment, the model which considers both visual and textual modalities as input (i.e. Image + TextDesc with OG + AO pair) is used, which is the best performing combination as reported in Table 6.3. As evident from Table 6.4, for different inference types, there are minor performance deviations when the prompt variations described above are used. For pre-condition, imperative prompt yields the best results. The model favors interrogative prompts to produce best inference for before and after events. The assertive (structured response) prompt is able to achieve the best performance for effect and goal inference. Based on the results, the use of assertive (sentence completion) prompts should be limited for all action-related inference dimensions.

Table 6.4 also provides insights related to how the model performs across each inference type. Across BLEU-2, METEOR, CIDER and Accuracy@50 metrics, the generated inference for pre-condition, before and after events is relatively better (with respect to the ground-truth) in comparison with effects and goals. Contrary to that, the uniqueness score for goals, effects and pre-conditions are higher. The novelty score for pre-conditions are significantly higher than the rest of the inference dimensions. However, low overall numbers throughout this table is indicating that state-of-the-art LLMs struggle to make concrete reasoning about actions, which humans can effortlessly perform in their day-to-day life.

Human Evaluation A human study is performed with 3 individuals with respect to 100 samples (20 samples for each inference type) from the dataset collected in this work. The evaluators were provided with an image, inference type of the interest and model generated natural language inference. They were advised to rate the sentence along two metrics:

- Correctness of the generated inference (choose from Incorrect / Partially-correct / Correct) i.e. provided the image and the desired inference type (e.g. effect), the participant indicates whether the generated inference was correct or not.
- Creativity of the model for the generated inference (choose from Obvious / Somewhat-Creative / Out-of-the-box) i.e. provided the image and inference type, the participant indicates whether the generated inference was apparent from the content of the image (obvious) or the model could generate something beyond what the image conveyed (out-of-the-box).

Cohen's kappa statistic is reported as a measurement of inter-annotator agreement (McHugh, 2012). The results are reported in Table 6.5, which indicates substantial to significant agreement across most inference dimensions.

In gist, an ability to reason about actions is crucial for Artificial Intelligent (AI) agents in closing the gap between human-level and machine performance in NLP, vision, and robotics tasks. In this work, an attempt is made to emphasize that beyond complex and multi-step physical interactions, actions have tightly coupled commonsense aspects associated with them, which include pre-conditions, executability, effects, high-level goals, and temporal dependency. A pipeline with multiple existing syntactic and semantic NLP tools is setup to automatically collect commonsense data from cooking videos. The problem is formulated as a vision-language task where provided an image, the system will generate text inferences around aforementioned

Inference (Prompt)	Correctness Kappa	Creativity Kappa
Pre-condition (Pp2)	0.78	0.76
Effect (Pe4)	0.68	0.66
Goals (Pg4)	0.74	0.83
Before event (Pb3)	0.81	0.81
After event (Pa3)	0.71	0.77

Table 6.5: Human Evaluation Results: Kappa Scores Denoting the Model's Ability to Generate Correct and Creative Inference About Actions Along Various Dimensions

aspects of actions, inspired by Park *et al.* (2020). From the ablations, it is demonstrated that integration between visual and textual commonsense reasoning is crucial in achieving the best performance. The experimental results indicate that there is still a large room for improvement in this research area and future research is encouraged for improved learning and reasoning about fine-grained commonsense. Systems equipped with such a capability will interact with humans in more natural ways, will have a better understanding of the world, and will be better partners for humans in performing day-to-day tasks. The code and links to the data are available at `https://github.com/shailaja183/ActionConceptLearning.git`.

Chapter 7

A ZERO-SHOT APPROACH FOR CONCEPT LEARNING ABOUT OBJECTS

7.1 Motivation

An ability to learn about new objects from a small amount of visual data and produce convincing linguistic justification about the presence/absence of certain concepts (that collectively compose the object) in novel scenarios is an important characteristic of human cognition. Even toddlers and young children are observed to perform such reasoning. For example, consider a toddler who observes different everyday objects in a new picture book and listens to their parents pronouncing the name of each object. Later on, when they encounter objects (observed in picture book) around their house, they can recall their names. Although the objects from picture book and around the house differ from each other in designs and colors, the kid is able to correctly identify the approporiate object. Such reasoning is enabled by continuous abstraction and categorization of objects encountered, which are two fundamental elements of human cognition Reinert *et al.* (2021).

In other words, objects are the building blocks of human's understanding of the world. In day-to-day lives, people constantly encounter diverse objects in their surroundings- by looking at images, reading textual descriptions, interacting with objects, or conversing with other humans. Human brain continuously categorizes the perceived information and develops mental representations of various objects by abstracting their properties Murphy (2010). For example, the mental representation of object 'chair' is formed using attributes such as an object that is having a backrest, legs, a seat, made of wood or plastic, etc. Similarly, an object 'bird' can be identi-

fied by the presence of a beak, feathers, legs, wings, etc. This allows the toddler to effortlessly identify novel kinds of chairs or birds.

The most common approach for distinguishing objects in Vision-Language (V&L) models is through supervised training over a set of labeled images. Despite remarkable advances in the field, the best existing systems still rely on a large number of annotated training examples. On the other hand, humans (including children) typically need only a handful of examples to robustly recognize the object and make meaningful generalizations Landau *et al.* (1988).

Another drawback of existing V&L models for object recognition is their black-box nature. These models rely on the similarity of visual features (obtained from the images) and possible class labels to distinguish between instances. As a result, the model's reasoning process behind a particular prediction is not explainable. Contrary to that, if a person is asked to identify a particular object, they would pay closer attention to the attributes (concepts) that an object is likely to have. Such a modular reasoning process allows humans to explain their decision linguistically. For example, because a particular bird has a red body, pointed crest, red-orange beak and black face (along with some more features), a person is likely to identify the bird as 'cardinal'.

Inspired by this aspect of human reasoning, in this work, a training-free concept learning strategy is proposed that is driven by large language models (LLMs). The hypothesis here is that LLMs trained on massive text corpora possess a vast amount of knowledge about various concepts. Such models can be queried to produce useful textual descriptions about visual appearance of an object. It can be followed by Visual Question Answering (VQA) systems to verify the presence/absence of LLM-generated concept descriptions in the unseen images. This will alleviate the reliance on visual recognition systems that are expensive in nature (from supervised data collection and training viewpoints) yet crucial for many V&L downstream tasks.

Contributions:

- A lightweight, interpretable framework for zero-shot visual concept learning is presented by combining an LLM with VQA systems.
- The effectiveness of the proposed approach over several fine-grained visual classification benchmarks is demonstrated in comparison with state-of-the-art methods.
- Relevant ablations are performed and a detailed analysis of the strengths and weaknesses of the proposed approach is discussed with respect to the CUB dataset Wah *et al.* (2011).

7.2 ZSVQA Task Overview

In this paper, the zero-shot visual question answering (ZS-VQA) task for visual concept learning (as shown in Figure 7.1) is investigated. For each object category c, a set of testing examples are provided. Each example consists of an image and a question pair, for which a model should predict an answer. Images are considered positive (I_{pos}) or negative (I_{neg}) depending on whether the category c present or not respectively. All images in I_{pos} would have the category c present, but images may demonstrate a broad range of view-variations (different pose/action, view angle, crop factor or relative position of c in the frame, etc.). Whereas, images in the negative set I_{neg} are about different object category c' such that $c' \neq c$. For each image in $I_{pos} \cup I_{neg}$ for a category c, a binary question Q of the form 'Is there a <c>?' is posed. For all images in I_{pos} set, the answer to the question Q would be 'Yes' (indicating the presence of object category c). Whereas for the images in set I_{neg}, the answer would be 'No' (indicating the absence of object category c).

There are three important parameters that distinguish various methods for visual concept learning using LLMs. The parameter values/ranges reported in this paper

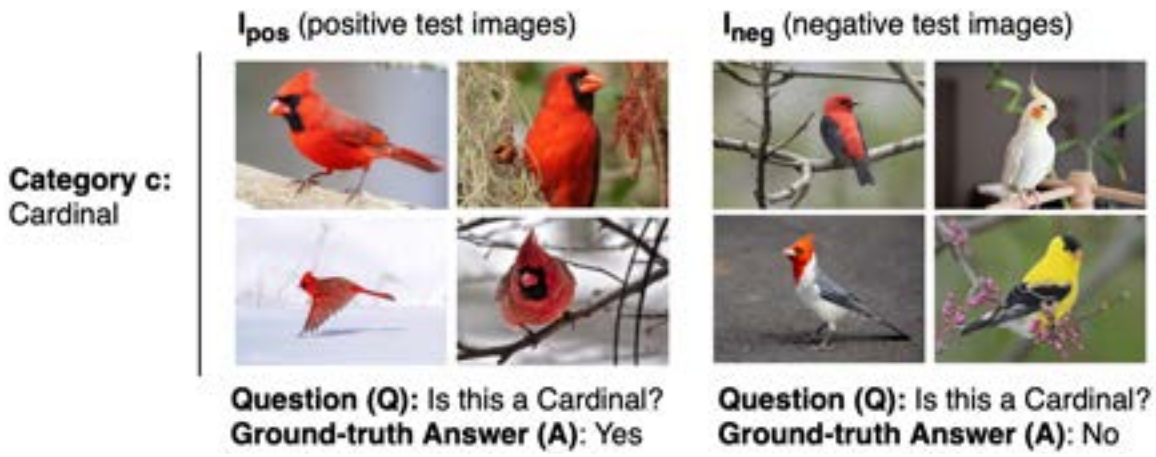


Figure 7.1: Zero-shot VQA Task Considered in This Paper Demonstrated Using 'Cardinal' Object Category from CUB (Wah *et al.*, 2011): Provided an Image and a Question as Input, the Model has to Predict 'Yes'/'No' Answer Depending on the Presence or Absence of the Object

are determined by respective authors (adapted from original manuscripts) based on their architecture design or empirical observations.

- **m** refers to the number of concepts generated to identify a particular object category c in the image and distinguish it from all other object categories $c' \neq c$.
- **n** refers to n-shot training or use of n labeled training examples for the preparation of a model; n=0 suggests that a given model is test-only or zero-shot (no training phase incorporated).
- **p** indicates whether the underlying method uses the fixed or dynamic prompt; Fixed prompt refers to the use of identical prompt for all object categories to obtain concept description using the LLM. Whereas dynamic prompt means a custom prompt is generated for every object category by leveraging prior knowledge (such as from external object taxonomy or additional information visible in the image).

7.3 Experiments

7.3.1 Proposed Model: Leveraging LLMs+VQA Systems for Interpretable Concept Learning

In this section, a zero-shot approach is described, which leverages knowledge of various concepts in large language model (GPT-3) and fine-grained visual understanding capabilities of visual question answering (VQA) systems. Figure 7.2 visually demonstrates the proposed approach- which can be divided into four key steps;

Step 1: Obtaining Concept Descriptors using GPT-3 Consider a test case from the CUB Wah *et al.* (2011)- fine-grained bird image classification dataset shown in Figure 7.2. The objective is to verify whether or not 'cardinal' is present in the image, as per the question. The first step is to query GPT-3 Brown *et al.* (2020) to retrieve concept descriptions. A template-based prompt is used of form 'Describe in m phrases separated by # – how the <object_category> looks like'. Where m is an odd integer and <object_category> is the object category that is to be verified in the image. More generally, as a result of this step, GPT-3 would generate m concept descriptions $\{v_1, v_2, .., v_m\}$ separated by #. For the example at hand, using m=3 and substituting <object_category> as 'cardinal', GPT-3 generates the response 'Bright red#Long Pointed Beak#Black Eyes'. These concept descriptions capture GPT-3's understanding about visual attributes of a cardinal.

The proposed system works with odd values of m as the majority is used as the aggregation mechanism in the later step. In this paper, experiments with m={1,3,5} are conducted and respective results are reported. Figure 7.3 demonstrates two examples from the CUB Wah *et al.* (2011) dataset when GPT-3 is prompted with different values of m. It can be observed that when m=1, the generated concept descriptions

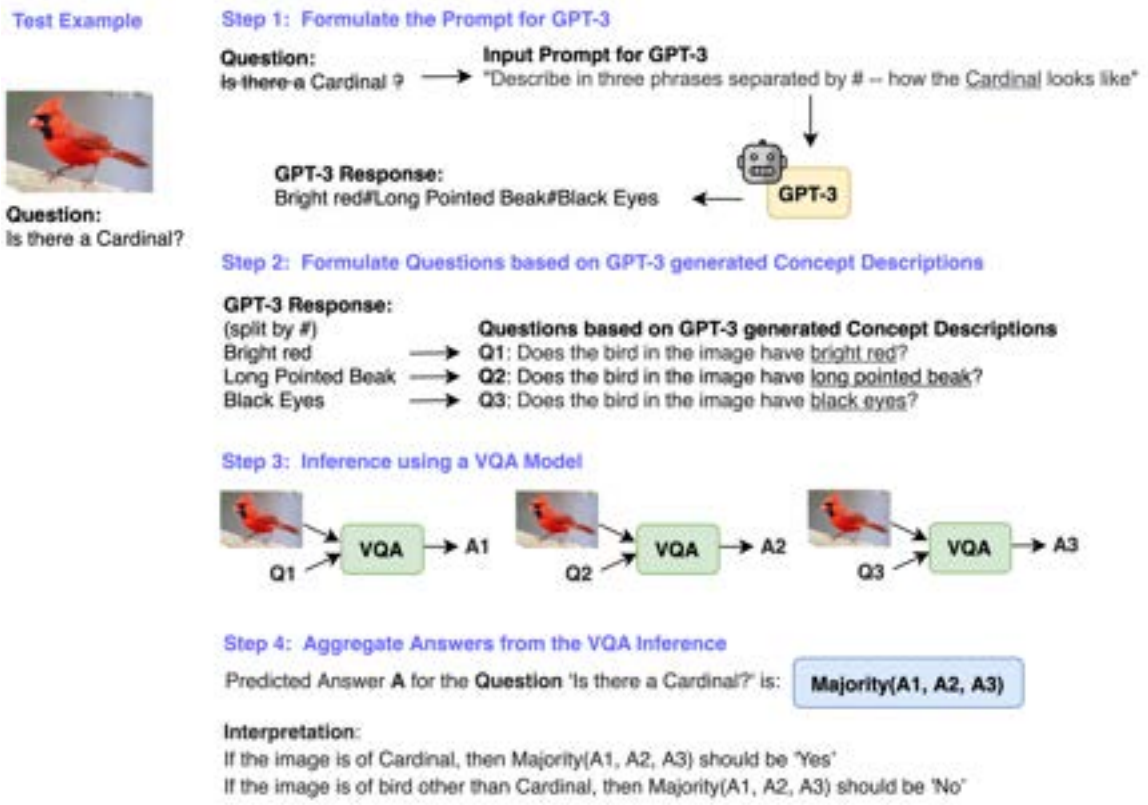


Figure 7.2: Overview of the Proposed Zero-shot Method That Uses LLM+VQA for Concept Learning: There are Four Key Steps- (i) Given an Object Category That Needs to be Verified in the Image, First GPT-3 is Queried Using a Predefined Prompt to Obtain the Concept Descriptions, (ii) Concept Descriptions Returned by GPT-3 are Turned Into a Set of Binary Meta-questions, (iii) Test Image Along With Each Meta-question is Posed to a VQA System, (iv) Aggregate Answers of All Meta-questions to Determine Presence or Absence of an Object Category in the Image

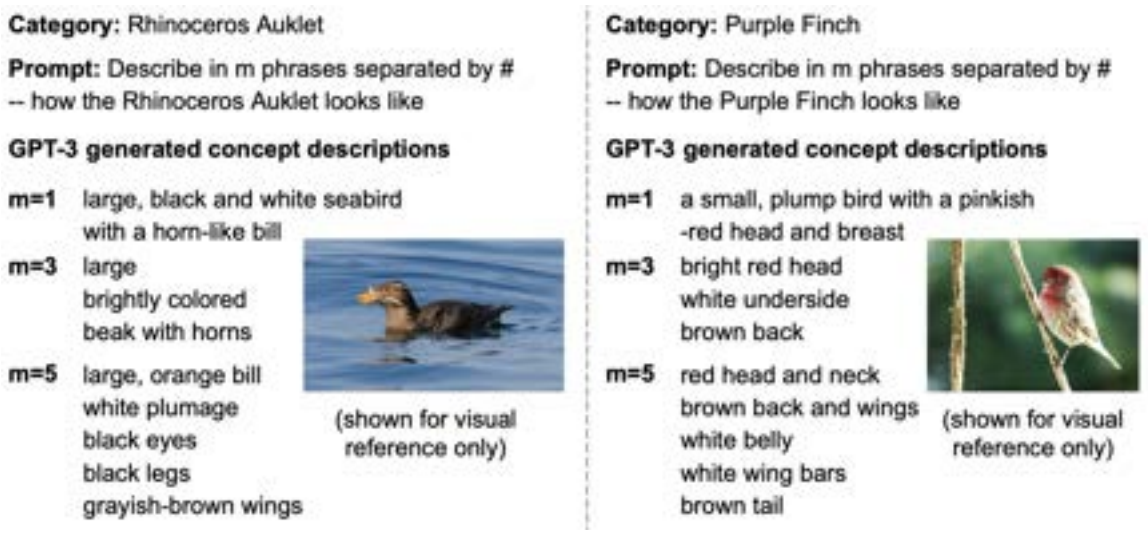


Figure 7.3: Two Categories from the CUB Dataset and Their Fine-grained Concept Descriptions Generated by GPT-3 for m={1,3,5}

are longer and typically includes a description of more than one body parts and its visual attributes. With m=3 and m=5, the generated phrases are short, concise, and discriminative from each other. A more detailed analysis of GPT-3 generated concept descriptions is provided in Section 7.4.

Step 2: Generating Binary Meta-questions The concept descriptions generated using GPT-3 (as described in the previous step) are in the form of phrases. For this task, an investigation is to be performed whether or not existing visual question answering (VQA) systems can verify the fine-grained visual descriptions generated by GPT-3 in the images. The VQA systems take questions as a linguistic input. Therefore, using a simple template 'Does the bird in the image have <v_i>?' (where i ranges from 1 to m), GPT-3 generated phrases are converted into a set of m binary questions. This is referred to as meta-questions corresponding to the concepts identified by GPT-3 for a given category of objects. For the cardinal example, three questions are formed: (1) Does the bird in the image have Bright red?, (2) Does the

bird in the image have Long Pointed Beak?, and (3) Does the bird in the image have Black Eyes?.

Step 3: Inference using VQA Models As mentioned earlier, the proposed method is primarily designed for zero-shot concept learning. The best off-the-shelf single-model architectures are adapted that support VQA tasks. Particularly, BLIP (Li *et al.*, 2022a), GIT (Wang *et al.*, 2022a) and ViLT (Kim *et al.*, 2021) are leveraged in this work. For each architecture, pre-trained version of the model is fine-tuned on the VQA dataset (Antol *et al.*, 2015a). These models are posed with the test image and meta-questions corresponding to the category (obtained from Step 2). In this case, the hypothesis is that- since these models are previously fine-tuned on the VQA dataset, they would be able to recognize different body parts of the bird and associated attributes (such as color, shape, size, texture).

Particularly, BLIP (Li *et al.*, 2022a) is a vision-language pre-training framework that effectively utilizes a synthetic caption generation along with a filter to identify noisy captions. It uses a multi-modal mixture of encoder-decoder architecture over large-scale noise-filtered image-caption data. It can transfer its learning well to both vision-language understanding and generation tasks in comparison with its predecessors. GIT (Wang *et al.*, 2022a) is a generative image-to-text transformer, which is simple network architecture (consisting of only one image encoder and one text decoder) that achieves strong performance with data-scaling. It can be used to perform a variety of vision-language tasks including visual question answering. Whereas, ViLT (Kim *et al.*, 2021) is a convolution-free pre-training approach that eliminates the need for obtaining object detection, object tags, OCR (optical character recognition), and region features from the image, which is faster, more parameter efficient yet demonstrates strong performance on downstream vision-language tasks.

Step 4: VQA Answers Aggregation Since each test image has m meta-questions corresponding, there are m binary (Yes/No) answers are generated in step 3. The final step is to aggregate the answers to conclude the presence or absence of the object category. For m=1, there is only one binary meta-question generated. For a positive example, the affirmative answer to the meta-question would determine the presence whereas for a negative example, the non-affirmative answer would indicate the absence of the concept. For m>1, a set of m answers will be generated by the VQA system therefore a majority voting is used as an aggregate measure for decision making. For this reason, taking odd values of m is recommended when working with the proposed model.

For m=3, if at least two answers among the three meta-questions for the positive and negative test cases are 'Yes' and 'No' respectively, then the prediction is considered correct. The threshold for experiments involving m=5 is set to three correct responses among the five meta-questions. More generally, in order to conclude the presence or absence of a particular category, the VQA model should be able to correctly answer at least (m+1)/2 meta-questions that are created based on the concept descriptions provided by GPT-3 for the respective object category.

7.4 Results and Discussion

7.4.1 Quantitative Results on CUB Dataset

In this section, results over the CUB dataset Wah *et al.* (2011) are reported and detailed analysis of the proposed method is presented including quality of GPT-3 Brown *et al.* (2020) generated concept descriptions, capability of existing VQA models to recognize fine-grained details in the images and comparison with existing concept learning methods (including zero-shot and one-shot approaches).

Zero-shot Performance The variations of the zero-shot concept recognition approach described above are prepared by using different VQA systems (BLIP, GIT, and ViLT) and experiments with a varied number of GPT-3 generated concepts (i.e. when m is 1/3/5) are performed. The proposed approach is compared with Menon and Vondrick (2022), (referred as Ext-CLIP) which is the only method that supports zero-shot evaluation. The performance with respect to the test partition of the CUB dataset Wah *et al.* (2011) is reported in Table 7.1. Additionally, there is a column corresponding to m=0 in the Table 7.1, which denotes the scenario where model does not take any LLM-generated concept descriptions into account. In other words, it can be thought of as having only a VQA model that takes test image and original test question (i.e. Is this a <category_name>?). This is intended to serve as a VQA baseline without any concept learning procedure.

	Accuracy(%) with varied value of m				
Method	0	1	3	5	var
Ours- BLIP	54.30	**70.98**	62.15	60.54	-
Ours - GIT	45.86	66.27	58.34	56.78	-
Ours- ViLT	48.33	67.12	59.05	53.28	-
Ext- CLIP	-	-	-	-	65.25

Table 7.1: Performance Comparison Between Variants of Proposed Zero-shot Concept Recognition Approach and Existing Work by Menon and Vondrick (2022) Denoted as Ext-CLIP over the CUB Test-set. Accuracy (%) is Used as an Evaluation Metric as the Underlying Task is Classification. 'Var' Denotes the use of Arbitrary Number of Concept Descriptions Produced by GPT-3 in the Ext-CLIP Model

All of the zero-shot model variations of proposed model with m=1 surpass the

performance of Ext-CLIP. There are three distinctions between this model and Ext-CLIP; First, both approaches use slightly different prompts to obtain concept descriptions from GPT-3. In this work, 'Describe in m phrases separated by # – how the <category_name> looks like' prompt is used whereas Ext-CLIP uses 'What are useful visual features for distinguishing a <category_name> in a photo?' prompt. Second, for a particular experiment, the value of m is pre-determined in proposed approach. Whereas Menon and Vondrick (2022) takes into account an arbitrary number of concept descriptions for each object category returned by GPT-3. Therefore the performance of Ext-CLIP is reported under the column 'var'. Finally, Ext-CLIP estimates CLIP similarity Radford *et al.* (2021) between images and sentence '<category_name> which is/has <concept_description>' at the test time. Contrary, a set of binary meta-questions are created in this work based on the concept descriptions which do not directly include <category_name>.

One-shot Performance There exists several methods in the literature Mei *et al.* (2021); Yang *et al.* (2023b); Cao *et al.* (2020) which are designed to learn new concepts from only one labelled example i.e. one-shot concept learning paradigm. Although the proposed method is designed primarily for the zero-shot concept learning, to compare the performance with existing methods, one-shot model is setup as follows: As evident from the Table 7.1, the proposed 4-step approach with BLIP as a VQA model performed the best. Therefore, a BLIP model is fine-tuned with the test image and one meta-question for each CUB category. In case the exists more than one meta-questions (m=3 and m=5 cases), one meta-question is randomly selected and used in the training. This one-shot fine-tuned BLIP model is leveraged in the Step 3 of the Section 7.3.1. All the remaining steps are identical to the zero-shot pipeline.

The results are summarized in Table 7.2 which compares the model in this work

with Ext- Falcon Mei *et al.* (2021), Ext- LaBo Yang *et al.* (2023b), and Ext- COMET Cao *et al.* (2020). Ext- LaBo Yang *et al.* (2023b) considers variable value of m in their model design (which are LLM generated concept descriptions) and Ext- COMET Cao *et al.* (2020) considers at most 15 concepts (which are human-curated for ground-truth) per object category (i.e. m¡=15). Ext- Falcon Mei *et al.* (2021) does not directly involve parameter m as their method attempts to learn visual concepts into box embeddings space. Since a meta-question is randomly sampled in case of m¿1, the reported results are averaged over three runs with a fixed seed value.

The best model BLIP$_{one-shot}$ (m=1), proposed in this work, yields the best accuracy 76.12%, which is 5.14% improvement over its zero-shot counterpart (reported in Table 7.1). The second best model is Ext- Falcon Mei *et al.* (2021), tailing the best performing model by ¡1% accuracy. The accuracy of the proposed BLIP$_{one-shot}$ model

Method	**Accuracy(%)**
Ours- BLIP$_{one-shot}$ (m=1)	**76.12**
Ours- BLIP$_{one-shot}$ (m=3)	69.46*
Ours- BLIP$_{one-shot}$ (m=5)	63.22*
Ext- Falcon	75.28
Ext- LaBo (m = var)	54.19
Ext- COMET (m¡=15)	67.9

Table 7.2: Accuracy(%) Comparison Between a Variant of Proposed One-shot Concept Recognition Approach and Existing Work by Mei *et al.* (2021); Yang *et al.* (2023b); Cao *et al.* (2020) Denoted as Ext- Falcon, Ext- LaBo, Ext- COMET Respectively on the CUB Test-set (* Indicates Averaged Results Over Three Runs)

variations decreased with the higher values of m (i.e. from 1 to 3 and 3 to 5), which is consistent to the observation in the zero-shot setting. Another interesting observation is that- for m=3, the performance gap between the zero-shot and one-shot model is maximum i.e. 7.31%. This indicates a possibility of generated meta-questions (for m=3 case) being more useful for the BLIP model to solve the downstream concept identification task (compared to m=1 and m=5). Although more experiments are needed to make appropriate conclusion, this is an interesting research question to explore in future.

Other existing models (Ext- LaBo and Ext- COMET), despite incorporating more number of LLM generated or human-curated concepts, underperformed compared to m=1 and m=3 model variations of the proposed model. Note that, among the models compared in Table 7.2, LaBo Yang *et al.* (2023b) and the proposed model are only two human-interpretable concept learning methods. Even though the Ext- Falcon Mei *et al.* (2021) achieves comparable accuracy to the proposed model, it lacks natural language explanations conveying model's decision of a particular object category.

Effect of the Choice of m on Performance The VQA task over the CUB dataset referred in this work is a binary VQA task i.e. whether a particular bird is present in the image or not. Therefore, the random chance accuracy for this task is 50%. The results from Table 7.1 indicate that VQA models that are not previously trained on the CUB data and directly tested over the bird identification task (m=0) perform almost close to the random chance. By simply replacing the category name with a single conceptual description generated by GPT-3 (i.e. m=1 case), the models gain 15-20% accuracy improvements. When the value of m increases further, a drop in the overall accuracy has been observed across all VQA variants. This trend is consistent with one-shot variants of the proposed model.

Two qualitative examples predicted by the best-performing zero-shot variant proposed in this paper are shown in Figure 7.4. The example on the top demonstrates incorporating multiple concise concept descriptions is beneficial. Accuracy for 'Eastern Towhee' category improves from 60.67% to 82% when m is set to 1 and 3 respectively. On the other hand, the example at the bottom demonstrates the case where a noisy visual attribute generated by GPT-3 (i.e. 'white eye ring') hurts the model performance (when m=3). As a result, accuracy for 'Vermilion Flycatcher' drops to 50.67% (m=3) compared to 97.33% (m=1). For both the success as well as the failure cases, the proposed model remains human-interpretable.

Moreover, an analyis is performed to understnad how the choice of m affects the individual object categories. To visualize this, in Figure 7.5, a plot on how the accuracy, number of False Positives (FP), and False Negatives (FN) for each category varies when m changes from 1 to 3. Accuracy denotes the number of correct test predictions using the zero-shot LLM+BLIP variant. Among 200 object categories in the CUB dataset, 120 categories demonstrate lower overall accuracy, and for 74

Figure 7.4: Two Qualitative Examples Predicted by the GPT-3+BLIP Model, Which is a Top-performing Zero-shot Variant in This Work

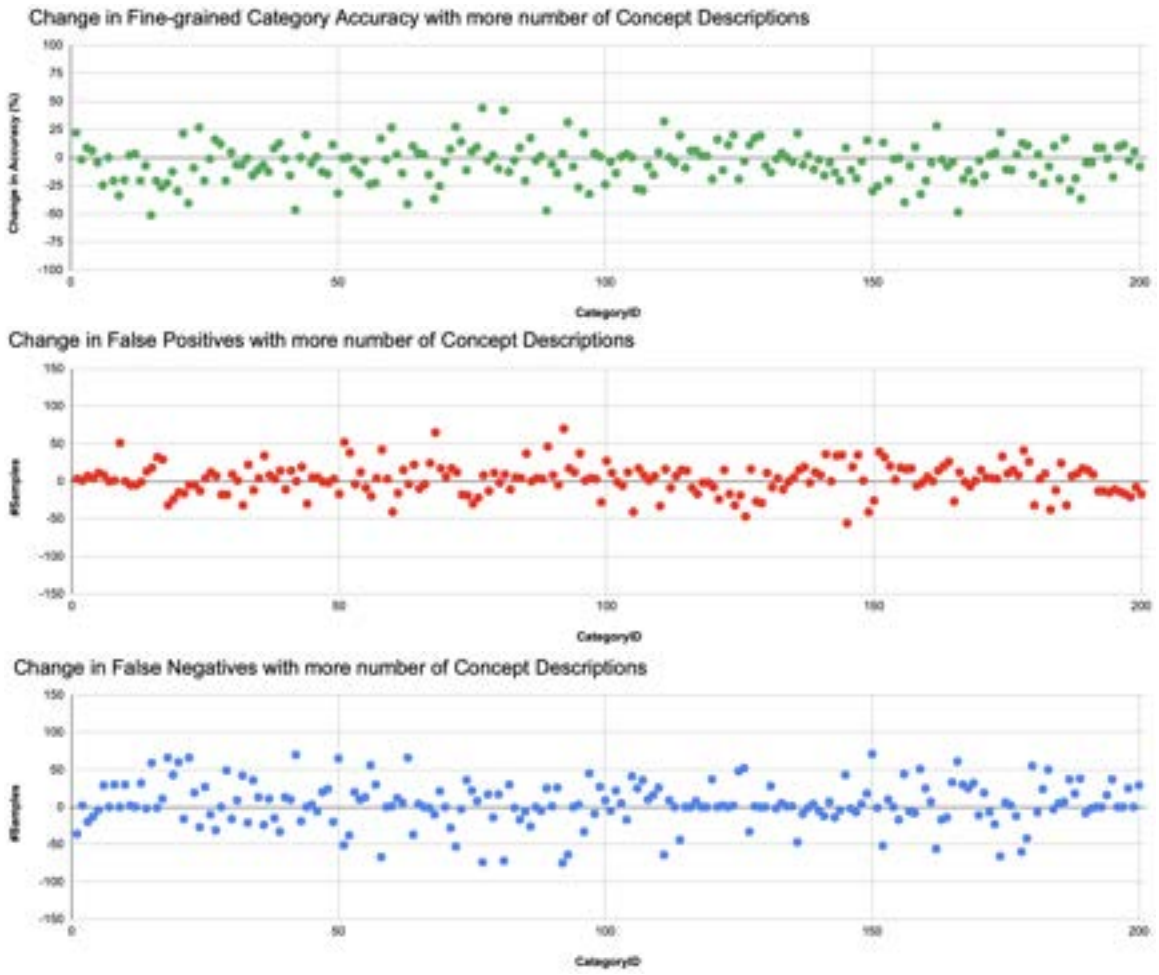


Figure 7.5: Plots Demonstrating How Accuracy (on Top), False Positives (in Mid), and False Negatives (at the Bottom) Per Object Category Change When More Number of Concept Descriptions are Obtained Using GPT-3 (i.e. When Changing m=1 to m=3) and Incorporated in the Best Zero-shot Model GPT-3+BLIP

categories the accuracy improves when m=3. Both accuracy gains and drops demonstrate 0-50% variation in comparison with accuracy obtained for the m=1 setting. For 73 and 78 categories, the count of FN and FP predictions are respectively lower compared to m=1.

7.4.2 Qualitative Results on CUB Dataset

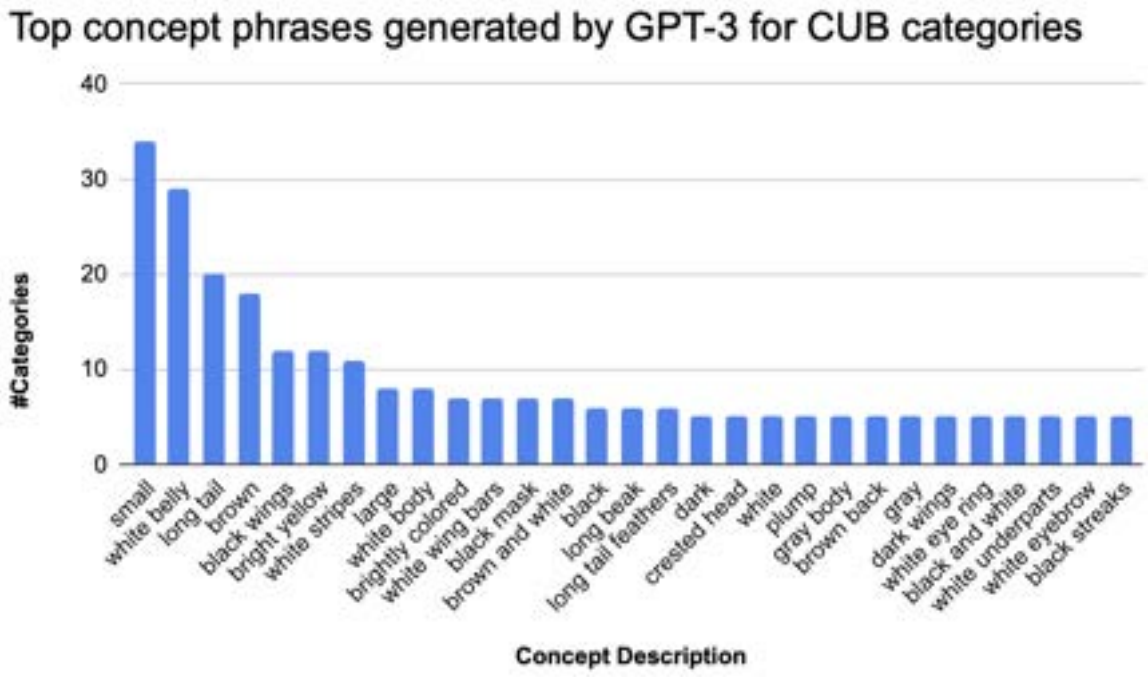


Figure 7.6: Most Frequent Concept Descriptions Returned by GPT-3 for 200 Object Categories in the CUB Dataset

Diversity of LLM Generated Concept Descriptions Below, an analysis is provided on how diverse are GPT-3 generated concept descriptors for all object categories in the CUB dataset. For 200 object categories, top-3 concept descriptions (m=3) are analyzed for each object category, among which 262 (∼42%) are unique. A plot of the most frequently generated descriptions for the CUB is shown in Figure 7.6. The phrases 'small' and 'white belly' are the most common across the dataset, specifically, shared across 34 and 29 categories respectively.

It can be observed that most concept descriptions generated by GPT-3 in the context of the CUB dataset include combinations of attributes such as color, shape, size, texture/pattern, and body parts. Therefore, 600 concept descriptions are manually

classified based on the combination of attributes reflected. For example, the description 'yellowish-olive' would be considered as 'Color', 'long tail' would be considered as 'Size+Body Part', and so on. The distribution based on The label 'Multiple' implies the presence of more than two attributes (e.g. 'brown streaked back' which includes color, texture/pattern, and body parts), The label 'Other' implies descriptions that do not fall into any of the standard combinations of attributes (e.g. 'three toes on each foot' was a concept description obtained for the category american three toed woodpecker). The distribution of concept descriptions based on the attribute types is shown in Figure 7.7. Among 600 descriptions analyzed, the attribute combinations 'Color+Body Parts' (e.g. 'yellow breast') and 'Color' (e.g. 'blue and white') cumulatively span 60% of the descriptions.

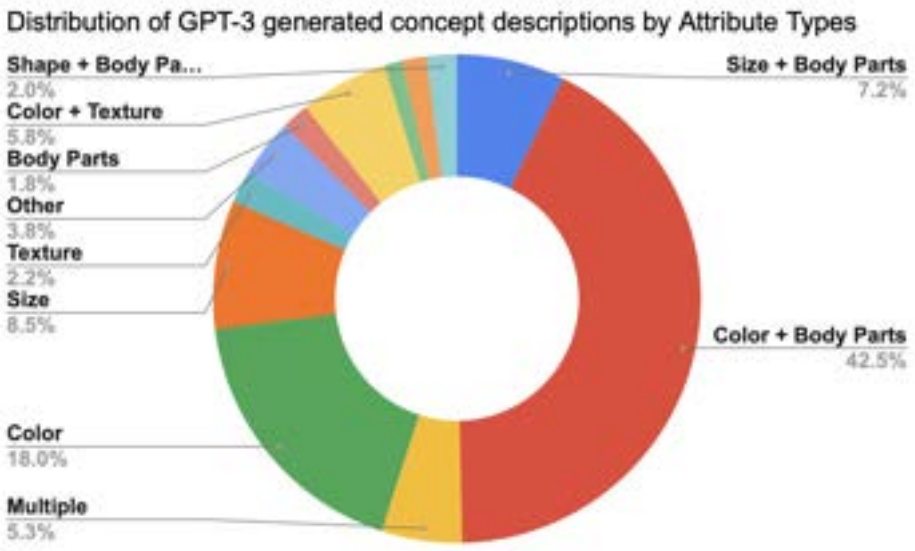


Figure 7.7: Distribution of GPT-3 Generated Concept Descriptions (m=3) for the CUB Dataset as Combinations of Attributes (Color, Shape, Size, Texture/Pattern, and Body Parts) Commonly Used to Identify Bird Species

Are VQA Systems Reliable for Fine-grained Attribute Recognition Described in Language? The proposed approach has two key components- an LLM and a VQA model. Hence, the end-task accuracy depends on- how correct are the LLM generated concept descriptions for various categories and how well the VQA system can verify the presence of those concept descriptions in the images. Assuming that GPT-3 has the perfect knowledge about appearance of various categories in the CUB dataset, in this section, an analysis is conducted on how reliable are existing VQA models for the task of fine-grained attribute recognition. Previously five major attribute types were identified i.e. colors, shapes, sizes, textures/patterns, body parts that are often used to describe CUB categories. For each attribute type, 3 examples are randomly selected from the dataset and shown in Table 7.3 and 7.4.

Two accuracy metrics are used for this evaluation- CategoryAccuracy and AttributeAccuracy. CategoryAccuracy denotes the correct number of answers generated by the proposed 4-stage model (GPT3+BLIP model with m=3) for all image-question pairs related to the given category in test set. On the other hand, AttributeAccuracy denotes the number of instances (within the test set for a given category) where the meta-question corresponding to the concept at hand (i.e. color 'yellowish-olive' or texture 'white spots on wings') is correctly verified by the model. Consider the example of 'western grebe' category tested for VQA model's ability to recognize 'Size' attribute (third row, first column in Table 7.3). The interpretation of CategoryAccuracy and AttributeAccuracy for this example can be justified as follows. Among all instances that aim to test 'western grebe' category in CUB's test data, 84.67% (i.e. CategoryAccuracy) of them can be correctly predicted by the proposed GPT3+LLM model. On the other hand, for all those examples, the BLIP model can answer the meta-question 'Does the bird in the image have long neck?' with accuracy 76% (i.e. AttributeAccuracy).

Qualitative analysis of VQA's ability to recognize various visual attribute types in CUB			
	Color		
Image:			
Category:	worm eating warbler	lazuli bunting	pied kingfisher
Concept:	yellowish-olive	bright blue	blue and white
CategoryAcc(%):	70.00	44.00	73.33
AttributeAccuracy(%):	97.33	97.33	30.67
	Shape		
Image:			
Category:	parakeet auklet	western meadowlark	white pelican
Concept:	round body	black v on chest	pouch-like bill
CategoryAcc(%):	49.33	52.67	91.33
AttributeAccuracy(%):	20.00	2.67	17.33
	Size		
Image:			
Category:	western grebe	anna hummingbird	grasshopper sparrow
Concept:	long neck	long beak	small body
CategoryAcc(%):	84.67	68.67	71.33
AttributeAccuracy(%):	76.00	81.33	100.00

Table 7.3: Qualitative Analysis of VQA System's Ability to Recognize Color, Shape, Size, Texture, Body Parts and Other Adjectives Generated by GPT-3 for Various CUB Object Categories

	Texture		
Image:			
Category:	palm warbler	groove billed ani	red cockaded woodpecker
Concept:	streaked	furry body	white spots on wings
CategoryAcc(%):	76.67	43.33	76.67
AttributeAccuracy(%):	80.00	70.66	85.33
	Body Parts		
Image:			
Category:	yellow headed blackbird	gray catbird	american 3-toed woodpecker
Concept:	white wing bars	white undertail	three toes on each foot
CategoryAcc(%):	96.67	38.00	48.67
AttributeAccuracy(%):	18.67	0.00	18.67
	Other Adjectives		
Image:			
Category:	common tern	cactus wren	scissor tailed flycatcher
Concept:	long forked tail	spiky feathers	bright colors
CategoryAcc(%):	73.33	78.00	60.67
AttributeAccuracy(%):	30.67	96.00	21.33

*

Table 7.4: Qualitative Analysis of VQA System's Ability to Recognize Color, Shape, Size, Texture, Body Parts and Other Adjectives Generated by GPT-3 for Various CUB Object Categories

Observing the results in Table 7.3 and 7.4, it can be concluded that attributes 'Color', 'Size', and 'Texture' are robustly recognized by the BLIP model. Whereas the AttributeAccuracy remains low for 'Shape' and 'Body Parts', indicative of model struggling with these attributes despite achieving high CategoryAccuracy for certain examples. This demonstrates the advantage of using higher values of m (such as 3 or 5) in the proposed model. In other words, by incorporating more than one fine-grained concept descriptions, it is possible to prevent single point of failures in the VQA stage of the model by relying on other visual attributes. This is especially the case where the generated concept description are hard to recognize in the image due to pose variations or insufficient/tiny details captured in the image. For example, consider the 'Scissor tailed flycatcher' category in Table 7.4 (bottom right corner). The concept generated by GPT-3 for this category is 'bright colors'. Due to lighting in this picture, it is hard to conclude whether this bird has bright colors or not. The BLIP model also struggles to recognize this attribute, evident from the low AttributeAccuracy i.e. 21.33%. However, the respective CategoryAccuracy is relatively high (60.67%), which indicates that model is possibly relying on other concept descriptions returned by GPT-3 and can correctly answer meta-questions based on them.

7.4.3 Quantitative Results for Other Visual Classification Datasets

To verify the generalization ability of the proposed LLM+VQA pipeline for concept learning about objects, in this section, further testing over existing image classification benchmarks is conducted. For this purpose, popular image classification datasets Krizhevsky *et al.* (2009); Bossard *et al.* (2014); Cimpoi *et al.* (2014) are leveraged. The characteristics of these datasets are briefly listed in Table 7.5. Programmatically these datasets are converted to the zero-shot VQA format described in Section 7.2.

Dataset	#Categories	Domain	Size
CIFAR-10	10	Mixed	10k
CIFAR-100	100	Mixed	10k
Food-101	101	Food	25.2k
DTD	47	Textures	1.8k

Table 7.5: Image Classification Datasets Evaluated in This Work with Relevant Information About Number of Distinct Categories Incorporated, Domain of the Images and Size of the Testing Data

The proposed method (in this work) is compared against four methods- LaBo (Yang *et al.*, 2023b), CDA (Yan *et al.*, 2023), CuPL (Pratt *et al.*, 2022) and VCD (Menon and Vondrick, 2022). All the aforementioned models use GPT-3 (Brown *et al.*, 2020) as a knowledge base to tackle visual concept learning task. LaBo extends the idea of concept bottleneck models proposed by Koh *et al.* (2020) with LLM-extracted concepts which supports few-shot evaluation. CuPL, CDA, VCD generate concept descriptions with LLMs and use them for knowledge-aware zero-shot image-text matching over CLIP (Radford *et al.*, 2021) to improve concept recognition. While the focus of Pratt *et al.* (2022) (referred to as CuPL) was on the automatic selection of the prompt to query the LLM, Yan *et al.* (2023) (referred to as CDA) distinguished themselves by identifying a small concise set of relevant attributes by avoiding noisy LLM generations and reduced computation.

The proposed method shares two-fold similarity with other approaches: first, zero-shot/few-shot evaluation of visual concept learning is supported, and second, all models support some level of interpretability. The proposed GPT3+VQA model relies on a fixed hand-crafted prompt like Menon and Vondrick (2022); Yang *et al.* (2023b) to

obtain visual descriptors for each object. However, object labels are not incorporated during the visual recognition step like Menon and Vondrick (2022). Unlike all foundation model-based methods that rely on CLIP for image-attribute matching, in this work, VQA models are explored to verify the presence or absence of attributes in the given image. The proposed approach yields comparable performance over existing methods considering only three concept descriptions per object (i.e. m=3), which is significantly lower than Yan *et al.* (2023) (i.e. m$\geq$16). Although there are many other image classifications datasets (not covered here), the choice of datasets is based on the popularity among the methods that are compared here. The only exception is the VCD model, which does not report performance on the CIFAR variants.

The results are reported in Table 7.6. As explained earlier, the choice of parameters m, n and p play a crucial role in the proposed method. The closest values of m, n and p parameters are reported for each method (to the best of knowledge and understanding of the architecture presented in the respective manuscript). Although due to the difference in these parameters, these methods cannot be compared directly, it is suggested to use the information in Table 7.6 to better understand the pros/cons of each method and their cumulative performance as a measure of how hard a particular image classification dataset is and to what extent LLMs are able to generate appropriate descriptions about the classes.

As observed in Table 7.6, most of the existing models for concept learning use fixed/templated prompt to obtain concept descriptions from the LLM. The only exception is the CuPL (Pratt *et al.*, 2022) model. Additionally, the proposed method, CuPL and VCD (Menon and Vondrick, 2022) are designed for zero-shot concept recognition. Whereas LaBo (Yang *et al.*, 2023b) requires at least 1 training example for each category. Finally, for the proposed model, CDA and VCD rely on a limited set of concept descriptions retrieved with the help of LLM. Contrary to that, LaBo

		Method (and corresponding Accuracy%)									
Dataset	**Configurable Parameters**	**LaBo**	**CDA**	**CuPL**	**VCD**	**Our (BLIP)**	**Our (BLIP)**	**Our (BLIP)**	**Our (ViLT)**	**Our (ViLT)**	**Our (ViLT)**
	m	50	4	1	var	1	3	5	1	3	5
	n	1	0	0	0	0	0	0	0	0	0
	p	fix	fix	var	fix	fix	fix	fix	fix	fix	fix
CIFAR-10		∼92%	∼67%	95.81%	-	61.42%	64.28%	51.37%	42.85%	62.81%	41.81%
CIFAR-100		∼63%	∼17%	78.47%	-	26.06%	33.73%	22.29%	27.11%	31.21%	16.27%
Food-101		∼80%	∼17%	93.33%	83.63%	82.95%	86.27%	84.30%	79.37%	79.77%	80.02%
DTD		∼53%	-	58.90%	44.26%	83.83%	88.94%	90.80%	87.87%	94.45%	96.12%

Table 7.6: Accuracy (%) Comparison Between Variants of the Proposed Zero-shot Concept Recognition Approach and Existing Work Related to Concept Learning Using LLMs over Multiple Image Classification Datasets. Parameters n, m and p Used in Each Method are Reported for Comparison. Approximate (∼) Symbol is Used When Exact Value is not Available

requires significantly large number of concept descriptions (50 descriptions per object category) to push the model performance. On the other hand, VCD dynamically determines the number of concept descriptions based on the example at hand (denoted as var).

The CuPL method achieves the best overall performance over CIFAR-10, CIFAR-100 and Food-101 datsets. All of the proposed model variants outperform other methods over DTD dataset, with notable performance gains. GPT3+ViLT combination with 5 concept descriptions yields the best performance overall (96.12%) for the DTD dataset. Moreover, the DTD dataset poses exception to the prior observation that performance boosts when changing m=1 to m=3 and declines when further increases to m=5. This possibly indicates that the concepts to be learnt (textures in this case) benefit from having multiple alternative descriptions that the model can

rely upon and avoid failure from limited attributes that may or may not be recognizable in the images. The proposed model demonstrates the second best performance over Food-101 dataset, tailing behind 7.06% by CuPL. LaBo achieves the comparable performance over the CIFAR-10 dataset in comparison with the best method CuPL. However, with more number of classes in CIFAR-100, the performance gap between CuPL and LaBo increases. The CDA consistently remains the low performing method across all datasets considered. The key takeaways from these experiments are summarized below:

- Concept learning methods that leverage LLMs work very well under zero-shot settings, which can be a promising alternative to expensive supervised training, which is expensive and time-consuming in nature.
- The construction of the prompt is a crucial step for all concept learning methods based on LLMs, The customized prompt formulation is likely to achieve superior results.
- In an ideal case, m>1 is desirable to avoid single point of failure in concept recognition pipeline, which is the unique aspect of our model proposed in this paper.

In summary, humans capitalize on concepts to learn mental representations of various objects and use this knowledge to identify novel instances of an object using only a few examples. Inspired by this, in this work, a framework for zero-shot visual concept learning is introduced. The linguistic knowledge about visual objects from large language models (LLM) is leveraged to generate concept descriptors. Instead of asking a VQA model whether or not a particular category is present in the image, the category identification process is broken down into multiple sub-questions based on the LLM generated concept descriptors. Using GPT-3 along with the state-of-the-art VQA models, this concept learning method demonstrates promising results over

the existing fine-grained image recognition datasets. The proposed method does not require significant computing overhead, improves performance over existing methods, and is human-interpretable to explain the model decisions. The code and links to the data are available at `https://github.com/shailaja183/ObjectConceptLearning.git`.

Chapter 8

CONCLUSION

This dissertation spans two prominent research directions in Artificial Intelligence (AI)- 'Multi-modal Understanding' and 'Reasoning about Actions and Change (RAC)'. A comprehensive survey of relevant works in both directions is conducted (Sampat *et al.*, 2024a, 2022b) and research problems are identified that are important from an applications perspective. Towards this end, new datasets are developed and made available for the community to pursue future research on these topics. Extensive experiments are performed using a wide variety of traditional and state-of-the-art neural models, neural-symbolic models, and large language models. Moreover, novel models are proposed that are more effective in terms of performance, data efficiency, and/or generalization capability. Detailed analysis of model performance is conducted and ablations are performed to emphasize the importance of various model components. By contributing to the development and evaluation of such models, an attempt is made to accomplish more robust and reliable AI systems that can perform complex reasoning and better collaborate with humans.

8.1 Summary of Contributions

8.1.1 Multi-modal Understanding

The importance of joint reasoning capability over image-text context is emphasized, inspired by how humans use such an ability in day-to-day tasks. To support the development of AI systems that can perform multi-modal reasoning, a Visuo-Linguistic Question Answering (VLQA) Dataset (Sampat *et al.*, 2020a) is proposed.

Specifically, provided visuals and a text passage, the task in VLQA is to answer questions by combining information from both reasoning over them, which are otherwise unanswerable. Experiments are conducted with state-of-the-art VQA models (including large-scale vision-language pre-training) to gauge their ability to perform multi-modal reasoning. An initial version of the VLQA dataset was observed to be relatively small and quite diverse, which is speculated to be a factor in the lower performance of data-driven models. In the subsequent work, a large-scale, multi-task visual-linguistic benchmark (VL-GLUE) (Sampat *et al.*, 2024a) is developed by augmenting, incorporating, and standardizing other multi-modal datasets (developed in parallel lines of work). The datasets in this benchmark are grouped into 7 clusters based on the complexity of image understanding and reasoning abilities required. By compiling this benchmark, systematic development of vision-language models is encouraged that have a broad range of multi-modal reasoning capabilities.

8.1.2 Reasoning About Actions and Change

Motivated by humans' ability to visualize numerous imaginary situations about a given visual scene and predict possible future outcomes, the CLEVR_HYP (Sampat *et al.*, 2021) dataset is developed. Particularly, this dataset evaluates the ability of VQA systems to predict effects that will be caused by performing hypothetical actions over the given image. The dataset is synthetically generated in a controlled environment, in order to focus on the models' reasoning ability without involving complex image recognition and open-ended language artifacts. Several baselines (including end-to-end neural, two-stage neural, and two-stage neural-symbolic) are developed based on existing vision and language architectures to tackle the hypothetical reasoning task. Follow-up work in this direction (Sampat *et al.*, 2022a) includes the development of an improved model for this task where two-fold action representation

learning is enforced; first, by observing changes in a pair of scene graphs (before and after the action is performed) and second, mapping the changes to corresponding linguistic action descriptions. Empirical results demonstrate that this two-fold learning strategy outperforms the baselines while being data-efficient and more generalizable. Finally, a case study on broader aspects of object and action concept learning (Sampat *et al.*, 2024b,c) using a large language model is conducted.

8.2 Potential Applications of This Work

8.2.1 Multi-modal Understanding

Deriving inference from combined visual and textual information is an important skill for humans to perform day-to-day tasks. For example, product assembly using instruction manuals, navigating roads while following street signs, interpreting visual representations (e.g., charts) in various documents such as newspapers and reports, understanding concepts using textbook-style learning, etc. A similar capability in AI systems is desirable where they can understand multi-modal information and generate inferences based on it. Particularly, AI-based document understanding systems and applications involving fast concept learning in low-resource settings are of key interest.

Document Understanding: There is an abundance of manuals, documents, newspapers, and books that contain texts and visuals that complement each other to aid readers' understanding of a given topic. In most cases, users interpret such visual-linguistic clues and make inferences beyond what is explicitly provided in such documents. For example, consider a bar chart representing the number of electoral votes obtained by two parties in an election, which are 306 and 232 respectively. Knowing that the number of electoral votes needed to win a clear majority is 270, one would infer that the first party won the election and the second party lacked 38 electoral

votes to secure a majority. Therefore, the development of AI systems that can analyze such multi-modal documents and uncover complex inferences would be useful in many applications including automated business analytics tools, in guiding policy development and data-driven decision-making.

Novel Concept Learning in Low-Resource Settings: Teaching methodologies often leverage the multi-modal understanding ability of students to teach about new concepts. In other words, to explain unfamiliar or novel concepts, teachers refer to concepts that students are already familiar with and provide additional clues to establish a mapping between the known concept and a new concept. For example, if a person wants to explain to another person what a 'zebra' looks like, who has previously seen a 'horse', they can do so by giving a description such as 'zebra is a horse-like animal with black-and-white stripes over the body'. Using such multi-modal clues, a person is able to aid basic understanding of the new concept 'zebra' without seeing it themselves. For example, a person can make conclusions such as a zebra would have four legs, or a zebra can walk/run. It would be interesting to explore- to what extent deep neural networks can perform such reasoning and whether it can reduce the need for training/fine-tuning or the amount of data required to robustly recognize novel concepts. This would have use cases for low-resource applications and autonomous systems deployed under constraints, where frequent training is not possible.

8.2.2 Reasoning About Actions and Change

Actions & change are tightly integrated with each other and are important aspects of humans' day-to-day life. Actions are fundamentally responsible for the dynamic nature of the world by causing changes. On the other hand, changes can be explained in terms of actions performed over the objects. Therefore, the AI system's ability

to visualize actions and reason about changes over visual and linguistic inputs is important in many applications. Particularly, robots deployed to perform on-demand tasks, in improving image or product search over the web, and as an alternative to physical prototyping are of key interest.

Robots Performing On-demand Tasks: An ability to reason about actions and change is essential for robots deployed to perform on-demand tasks (whether it is homes, workplaces, or even battlefields). When a user instructs the robot to perform a particular task, it should perform hypothetical reasoning about the feasibility and consequences of performing the actions at a high level. If concluded to be viable for execution, the robot should decide on a course of action and deal with any undesirable situations that may arise.

Image Retrieval or Product Search Over Web: In a typical image search over the web, a user has a mental representation of the kinds of images they are looking for. Then the user expresses concepts related to that mental representation in language to query the search engine. The search normally begins with simple keywords and the description is iteratively refined until the image satisfying the desired criteria is obtained. This process is often time-consuming and in many cases leads to noisy results, which can be frustrating on the user end. An AI system that can perform hypothetical reasoning can be integrated with existing image retrieval algorithms to ensure the quality of the search results.

After a long browsing session, users often tend to forget the exact product name and keywords used in the past to retrieve the desired product or misplace the web links. They only remember vague details about the brand and visual details about how the product looked like. In such cases, a search strategy can be developed where

users start with similar-looking images from the web as a reference, and then express the changes or modifications as a text/instruction that would bring it closer to their mental representation of the product.

Rapid Prototyping: The designers or manufacturers often have new ideas or variations related to their product offerings. In most cases, they rely on prototypes or working models before production at scale. However, prototyping for all possible variations of a product might not be viable and can be quite expensive for many domains (cars, jewelry, etc.). A similar situation arises when customers would like personalized products but are unsure of how the actual product would look or whether or not it will suit their personality. In such cases, a model with hypothetical reasoning abilities with respect to images might be a handy, fast, and cost-effective solution.

8.3 Takeaways and Future Research Avenues

8.3.1 Multi-modal Understanding

Multi-modal learning aims to build models that can process and relate information from two or more modalities. Image-text multi-modality has received significant interest in the AI community recently as it is an important skill for humans to perform day-to-day tasks. Perception systems can be leveraged to identify a variety of visual information and a concise way to learn through observations. On the other hand, language provides an effective way to exchange thoughts, communicate, query, or provide explanations about a particular situation. As a result, a combination of visual and textual modalities provides natural flexibility for varied inference tasks that can reflect the reasoning required to navigate real-world scenarios. On the downside, the presence of multiple kinds of inputs makes the reasoning process complicated as information is now spanned across them and requires cross-inferencing. There has

been a growing interest in the development of systems that can reason about multi-modal contexts, however, a majority of them indicate a significant gap in comparison with human performance on this task.

Multi-modality is Beyond Image-Text Similarity: For most vision-language tasks, the role of textual components has been limited to evaluation purposes rather than as a modality that provides contextual information. This flaw in the design of the existing vision-language tasks has encouraged pre-training approaches that highly rely on image-text similarity to learn representations. In other words, by having image captioning datasets as a major source for pre-training, models capture representations that are common to both images and the respective text. Hence, models trained in this fashion struggle to perform joint inference over inputs where images and texts convey non-overlapping information. There is a need for the development of better pre-training strategies and appropriate training data which enables models to generate inferences over heterogeneous modalities. This would be crucial in improving the multi-modal reasoning capabilities of existing AI systems.

8.3.2 Reasoning About Actions and Change

Intelligent systems that interact with humans and assist them in performing tasks require two key abilities- understanding the environment as a combination of visual and linguistic modalities and performing relevant actions to achieve goals. Often, such systems operate in open and highly uncertain environments for which having a physical commonsense, understanding of goal-driven action-effects, and performing hypothetical reasoning are essential. Reasoning about actions is a relatively new research direction for the vision+language community; although there exists an array of well-defined tasks and datasets to start with. Though existing approaches can

achieve reasonably good performance on carefully crafted benchmarks, they lack a consolidated understanding of actions that is necessary to demonstrate robustness and generalization to be useful in real-world applications. There are two key recommendations in this regard;

A Synergy Between Synthetic and Realistic Benchmarks for Action Reasoning: There exists an array of synthetic and realistic datasets that tackle a broad range of reasoning about actions. Often, they differ from each other by considerations made during dataset creation e.g. kinds of actions and reasoning incorporated, the source of data, choice of task (generative vs. classification), and data format. This aspect of the existing datasets limits model evaluation on multiple relevant benchmarks (supervised models specifically) and ignores the transferability of the learned model to adapt to tasks in a real-world setting. The future development of benchmarks should be encouraged to have both synthetic and realistic components with common assumptions. So that, it is possible to train the model on large-scale synthetic benchmarks without picking up spurious correlations and yet being explainable. Later, it can be tested for generalization on the realistic counterpart of the data without expensive fine-tuning or in a few-shot setting. Building multi-task models is another workaround for this challenge, which is well explored in recent works. Multi-task learning has been shown to work well for tasks that share significant commonalities, however, it is not guaranteed to work for completely unrelated tasks.

Unified Action Understanding: Humans have consolidated physical, spatial, temporal, visual, social, and causal understanding of actions. Most existing benchmarks for action reasoning incorporate only one or two aforementioned aspects for the simplicity and focused model evaluation process. As a result, a model trained on

a particular benchmark lacks the big picture and fails to understand complex interactions with other aspects that were not observed during training. Davis and Marcus (2015) emphasize that humans can reason about situations of a kind 'if you slice a finger while preparing a meal, your prospective parents-in-law may not be impressed'. Though existing models seem to understand actions at a high level in isolation, reasoning about how they interrelate (as in the above example) is substantially unsolved. The best models are still dataset-specific and lack a consolidated understanding of actions that are necessary to demonstrate robustness and generalization in real-world applications. In this regard, there is a need for an umbrella task that supports end-to-end understanding of actions- from understanding action intents and pre-conditions to imagining outcomes of hypothetical situations and demonstrating social commonsense.

REFERENCES

Acharya, M., K. Kafle and C. Kanan, "Tallyqa: Answering complex counting questions", in "Proceedings of the AAAI Conference on Artificial Intelligence", vol. 33 (2019).

Adesam, Y., A. Berdicevskis and F. Morger, "Swedishglue–towards a swedish test set for evaluating natural language understanding models", (2020).

Andreas, J., M. Rohrbach, T. Darrell and D. Klein, "Neural module networks", in "2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016", pp. 39–48 (IEEE Computer Society, 2016), URL `https://doi.org/10.1109/CVPR.2016.12`.

Antol, S., A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick and D. Parikh, "Vqa: Visual question answering", in "Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)", pp. 2425–2433 (2015a).

Antol, S., A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick and D. Parikh, "VQA: visual question answering", in "2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015", pp. 2425–2433 (IEEE Computer Society, 2015b), URL `https://doi.org/10.1109/ICCV.2015.279`.

Auer, S., C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak and Z. Ives, "Dbpedia: A nucleus for a web of open data", in "The Semantic Web: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007+ ASWC 2007, Busan, Korea, November 11-15, 2007. Proceedings", pp. 722–735 (Springer, 2007).

Baral, C., "Reasoning about actions and change: from single agent actions to multi-agent actions", in "Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning (KR)", (2010).

Battaglia, P. W., J. B. Hamrick and J. B. Tenenbaum, "Simulation as an engine of physical scene understanding", Proceedings of the National Academy of Sciences **110**, 45, 18327–18332 (2013).

Ben Abacha, A., S. A. Hasan, V. V. Datla, D. Demner-Fushman and H. Müller, "Vqa-med: Overview of the medical visual question answering task at imageclef 2019", in "Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes", (9-12 September 2019, 2019).

Bhagavatula, C., R. Le Bras, C. Malaviya, K. Sakaguchi, A. Holtzman, H. Rashkin, D. Downey, W.-t. Yih and Y. Choi, "Abductive commonsense reasoning", in "International Conference on Learning Representations", (2019).

Bian, N., X. Han, L. Sun, H. Lin, Y. Lu and B. He, "Chatgpt is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models", (2023).

Bisk, Y., R. Zellers, J. Gao, Y. Choi *et al.*, "Piqa: Reasoning about physical commonsense in natural language", in "Proceedings of the AAAI conference on artificial intelligence", vol. 34, pp. 7432–7439 (2020a).

Bisk, Y., R. Zellers, J. Gao, Y. Choi *et al.*, "Piqa: Reasoning about physical commonsense in natural language", in "Proceedings of the AAAI Conference on Artificial Intelligence", (2020b).

Biten, A. F., R. Tito, A. Mafla, L. Gomez, M. Rusinol, E. Valveny, C. Jawahar and D. Karatzas, "Scene text visual question answering", in "Proceedings of the IEEE/CVF international conference on computer vision", pp. 4291–4301 (2019).

Blender Online Community, *Blender - a 3D modelling and rendering package*, Blender Foundation, URL `http://www.blender.org` (2019).

Bossard, L., M. Guillaumin and L. Van Gool, "Food-101–mining discriminative components with random forests", in "Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13", pp. 446–461 (Springer, 2014).

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners", Advances in neural information processing systems **33**, 1877–1901 (2020).

Cao, K., M. Brbic and J. Leskovec, "Concept learners for few-shot learning", in "International Conference on Learning Representations", (2020).

Cañete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang and J. Pérez, "Spanish pre-trained bert model and evaluation data", in "PML4DC at ICLR 2020", (2020).

Central Intelligence Agency, D., Washington, "The world factbook", (2019).

Chalkidis, I., A. Jana, D. Hartung, M. Bommarito, I. Androutsopoulos, D. Katz and N. Aletras, "Lexglue: A benchmark dataset for legal language understanding in english", in "Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)", pp. 4310–4330 (2022).

Chang, Y., M. Narang, H. Suzuki, G. Cao, J. Gao and Y. Bisk, "Webqa: Multihop and multimodal qa", in "Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition", pp. 16495–16504 (2022).

Chattopadhyay, P., R. Vedantam, R. R. Selvaraju, D. Batra and D. Parikh, "Counting everyday objects in everyday scenes", in "Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition", (2017).

Chen, L., G. Lin, S. Wang and Q. Wu, "Graph edit distance reward: Learning to edit scene graph", in "European Conference on Computer Vision", pp. 539–554 (Springer, 2020a), URL `https://www.ecva.net/papers/eccv_2020/papers_ECCV/papers/123640528.pdf`.

Chen, W., H. Zha, Z. Chen, W. Xiong, H. Wang and W. Y. Wang, "Hybridqa: A dataset of multi-hop question answering over tabular and textual data", in "Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings", pp. 1026–1036 (2020b).

Chen, X., H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár and C. L. Zitnick, "Ms coco captions: Data collection and evaluation server", arXiv preprint arXiv:1504.00325 (2015).

Chen, Y., Y. Liu, L. Dong, S. Wang, C. Zhu, M. Zeng and Y. Zhang, "AdaPrompt: Adaptive model training for prompt-based NLP", in "Findings of the Association for Computational Linguistics: Empirical Methods in Natural Language Processing 2022", pp. 6057–6068 (Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022), URL `https://aclanthology.org/2022.findings-emnlp.448`.

Cimpoi, M., S. Maji, I. Kokkinos, S. Mohamed and A. Vedaldi, "Describing textures in the wild", in "Proceedings of the IEEE conference on computer vision and pattern recognition", pp. 3606–3613 (2014).

Clark, P., I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick and O. Tafjord, "Think you have solved question answering? try arc, the ai2 reasoning challenge", arXiv preprint arXiv:1803.05457 (2018).

Dalvi, B., N. Tandon, A. Bosselut, W.-t. Yih and P. Clark, "Everything happens for a reason: Discovering the purpose of actions in procedural text", in "Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)", (2019).

Davis, E. and G. Marcus, "Commonsense reasoning and commonsense knowledge in artificial intelligence", In Communications of ACM (2015).

Denkowski, M. and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language", in "Proceedings of the ninth workshop on statistical machine translation", pp. 376–380 (2014).

Feng, L., J. Yu, D. Cai, S. Liu, H. Zheng and Y. Wang, "Asr-glue: A new multi-task benchmark for asr-robust natural language understanding", arXiv preprint arXiv:2108.13048 (2021).

Gao, Q., S. Yang, J. Chai and L. Vanderwende, "What action causes this? towards naive physical action-effect prediction", in "Proceedings of the Annual Meeting of the Association for Computational Linguistics", (2018).

Geman, D., S. Geman, N. Hallonquist and L. Younes, "Visual turing test for computer vision systems", Proceedings of the National Academy of Sciences **112**, 12, 3618–3623 (2015).

Gokhale, T., S. Sampat, Z. Fang, Y. Yang and C. Baral, "Cooking with blocks: A recipe for visual reasoning on image-pairs", in "Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops", pp. 5–8 (2019).

Gomes, L., "Machine-learning maestro michael jordan on the delusions of big data and other huge engineering efforts", URL https://tinyurl.com/yn9sw3sn (2014).

Gordon, D., A. avi, M. Rastegari, J. Redmon, D. Fox and A. Farhadi, "IQA: visual question answering in interactive environments", in "2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018", pp. 4089–4098 (IEEE Computer Society, 2018), URL http://openaccess.thecvf.com/content_cvpr_2018/html/Gordon_IQA_Visual_Question_CVPR_2018_paper.html.

Goyal, Y., T. Khot, D. Summers-Stay, D. Batra and D. Parikh, "Making the v in vqa matter: Elevating the role of image understanding in visual question answering", in "roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition", pp. 6904–6913 (2017).

Gupta, A., P. Dollar and R. Girshick, "Lvis: A dataset for large vocabulary instance segmentation", in "Proceedings of the IEEE/CVF conference on computer vision and pattern recognition", pp. 5356–5364 (2019).

Gurari, D., Q. Li, C. Lin, Y. Zhao, A. Guo, A. Stangl and J. 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", pp. 939–948 (2019).

He, K., G. Gkioxari, P. Dollár and R. B. Girshick, "Mask R-CNN", in "IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017", pp. 2980–2988 (IEEE Computer Society, 2017), URL https://doi.org/10.1109/ICCV.2017.322.

He, K., X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition", in "2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016", pp. 770–778 (IEEE Computer Society, 2016), URL https://doi.org/10.1109/CVPR.2016.90.

He, X., Q. H. Tran, G. Haffari, W. Chang, Z. Lin, T. Bui, F. Dernoncourt and N. Dam, "Scene graph modification based on natural language commands", in "Findings of the Association for Computational Linguistics: Empirical Methods in Natural Language Processing 2020", pp. 972–990 (Association for Computational Linguistics, 2020), URL https://www.aclweb.org/anthology/2020.findings-emnlp.87.

Hendrycks, D., C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song and J. Steinhardt, "Measuring massive multitask language understanding", in "International Conference on Learning Representations", (2020).

Higgs, L., "Gas price storm?", The Star-Ledger URL https://in.pinterest.com/pin/300896818868315676/ (2016).

Holtzman, A., J. Buys, L. Du, M. Forbes and Y. Choi, "The curious case of neural text degeneration", in "International Conference on Learning Representations", (2019).

Honda, "Honda: 2022 civic sedan owner's manual", Honda: 2022 Civic Sedan Owner's Manual URL https://techinfo.honda.com/rjanisis/pubs/OM/AH/AT2022220M/enu/AT2022220M.PDF (2022).

Honnibal, M., I. Montani, S. Van Landeghem, A. Boyd *et al.*, "spacy: Industrial-strength natural language processing in python", https://spacy.io (2020).

Hu, H., I. Misra and L. van der Maaten, "Evaluating text-to-image matching using binary image selection (bison)", in "The IEEE International Conference on Computer Vision (ICCV) Workshops", (2019).

Hu, J., S. Ruder, A. Siddhant, G. Neubig, O. Firat and M. Johnson, "Xtreme: a massively multilingual multi-task benchmark for evaluating cross-lingual generalization", in "Proceedings of the 37th International Conference on Machine Learning", pp. 4411–4421 (2020).

Huang, Z., Z. Zeng, Y. Huang, B. Liu, D. Fu and J. Fu, "Seeing out of the box: End-to-end pre-training for vision-language representation learning", in "Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition", pp. 12976–12985 (2021).

Hudson, D. A. and C. D. Manning, "Gqa: A new dataset for real-world visual reasoning and compositional question answering", in "Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)", pp. 6700–6709 (2019), URL https://openaccess.thecvf.com/content_CVPR_2019/papers/Hudson_GQA_A_New_Dataset_for_Real-World_Visual_Reasoning_and_Compositional_CVPR_2019_paper.pdf.

Isola, P., J. Lim and E. Adelson, "Discovering states and transformations in image collections", in "Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition", (2015).

Iyer, S., N. Dandekar and K. Csernai, "First quora dataset release: Question pairs", data. quora. com URL https://www.kaggle.com/c/quora-question-pairs/data (2017).

Jiang, Z., F. F. Xu, J. Araki and G. Neubig, "How can we know what language models know?", Transactions of the Association for Computational Linguistics **8**, 423–438 (2020).

Johnson, J., B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick and R. B. Girshick, "CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning", in "2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017", pp. 1988–1997 (IEEE Computer Society, 2017a), URL https://doi.org/10.1109/CVPR.2017.215.

Johnson, J., B. Hariharan, L. van der Maaten, J. Hoffman, L. Fei-Fei, C. L. Zitnick and R. B. Girshick, "Inferring and executing programs for visual reasoning", in "IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017", pp. 3008–3017 (IEEE Computer Society, 2017b), URL https://doi.org/10.1109/ICCV.2017.325.

Juba, B., "Integrated common sense learning and planning in pomdps", Journal of Machine Learning Research **17**, 96, 1–37, URL http://jmlr.org/papers/v17/13-584.html (2016).

Kafle, K., B. L. Price, S. Cohen and C. Kanan, "DVQA: understanding data visualizations via question answering", in "2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018", pp. 5648–5656 (IEEE Computer Society, 2018), URL http://openaccess.thecvf.com/content_cvpr_2018/html/Kafle_DVQA_Understanding_Data_CVPR_2018_paper.html.

Kahou, S. E., V. Michalski, A. Atkinson, Á. Kádár, A. Trischler and Y. Bengio, "Figureqa: An annotated figure dataset for visual reasoning", arXiv preprint arXiv:1710.07300 URL https://openreview.net/pdf?id=SyunbfbAb (2017).

Kakwani, D., A. Kunchukuttan, S. Golla, N. Gokul, A. Bhattacharyya, M. M. Khapra and P. Kumar, "Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages", in "Findings of the Association for Computational Linguistics: EMNLP 2020", pp. 4948–4961 (2020).

Kembhavi, A., M. Salvato, E. Kolve, M. Seo, H. Hajishirzi and A. Farhadi, "A diagram is worth a dozen images", in "Proceedings of the European Conference on Computer Vision (ECCV)", (2016).

Kembhavi, A., M. Seo, D. Schwenk, J. Choi, A. Farhadi and H. Hajishirzi, "Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension", in "Proceedings of the IEEE Conference on Computer Vision and Pattern recognition", pp. 4999–5007 (2017).

Khanuja, S., S. Dandapat, A. Srinivasan, S. Sitaram and M. Choudhury, "Gluecos: An evaluation benchmark for code-switched nlp", in "Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics", pp. 3575–3585 (2020).

Khot, T., A. Sabharwal and P. Clark, "SciTail: A textual entailment dataset from science question answering", in "Proceedings of the AAAI Conference on Artificial Intelligence", (2018).

Kim, W., B. Son and I. Kim, "Vilt: Vision-and-language transformer without convolution or region supervision", in "International Conference on Machine Learning", pp. 5583–5594 (PMLR, 2021).

Koh, P. W., T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim and P. Liang, "Concept bottleneck models", in "International conference on machine learning", pp. 5338–5348 (PMLR, 2020).

Koto, F., A. Rahimi, J. H. Lau and T. Baldwin, "Indolem and indobert: A benchmark dataset and pre-trained language model for indonesian nlp", in "Proceedings of the 28th International Conference on Computational Linguistics", pp. 757–770 (2020).

Kottur, S., J. M. F. Moura, D. Parikh, D. Batra and M. Rohrbach, "CLEVR-dialog: A diagnostic dataset for multi-round reasoning in visual dialog", in "Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)", pp. 582–595 (Association for Computational Linguistics, 2019), URL `https://www.aclweb.org/anthology/N19-1058`.

Koupaee, M. and W. Y. Wang, "Wikihow: A large scale text summarization dataset", arXiv preprint arXiv:1810.09305 (2018).

Krizhevsky, A., G. Hinton *et al.*, "Learning multiple layers of features from tiny images", (2009).

Kurihara, K., D. Kawahara and T. Shibata, "Jglue: Japanese general language understanding evaluation", in "Proceedings of the Thirteenth Language Resources and Evaluation Conference", pp. 2957–2966 (2022).

Lai, G., Q. Xie, H. Liu, Y. Yang and E. Hovy, "RACE: Large-scale ReAding comprehension dataset from examinations", in "Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing", pp. 785–794 (Association for Computational Linguistics, 2017), URL `https://www.aclweb.org/anthology/D17-1082`.

Lan, Z., M. Chen, S. Goodman, K. Gimpel, P. Sharma and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations", in "International Conference on Learning Representations", (2019).

Landau, B., L. B. Smith and S. S. Jones, "The importance of shape in early lexical learning", Cognitive development **3**, 3, 299–321 (1988).

Lau, J. J., S. Gayen, A. Ben Abacha and D. Demner-Fushman, "A dataset of clinically generated visual questions and answers about radiology images", Scientific data **5**, 1, 1–10 (2018).

Le, H., L. Vial, J. Frej, V. Segonne, M. Coavoux, B. Lecouteux, A. Allauzen, B. Crabbé, L. Besacier and D. Schwab, "Flaubert: Unsupervised language model pre-training for french", in "Proceedings of the Twelfth Language Resources and Evaluation Conference", pp. 2479–2490 (2020).

LeCun, Y., "Some obvious facts related to human level ai", URL `https://twitter.com/ylecun/status/1526672565233758213` (2022).

Li, D., J. Li, H. Le, G. Wang, S. Savarese and S. C. Hoi, "LAVIS: A one-stop library for language-vision intelligence", in "Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)", pp. 31–41 (Association for Computational Linguistics, Toronto, Canada, 2023), URL https://aclanthology.org/2023.acl-demo.3.

Li, J., D. Li, C. Xiong and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation", in "International Conference on Machine Learning", pp. 12888–12900 (PMLR, 2022a).

Li, L., J. Lei, Z. Gan, L. Yu, Y.-C. Chen, R. Pillai, Y. Cheng, L. Zhou, X. E. Wang, W. Y. Wang *et al.*, "Value: A multi-task benchmark for video-and-language understanding evaluation", in "Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)", (2021).

Li, L. H., M. Yatskar, D. Yin, C.-J. Hsieh and K.-W. Chang, "Visualbert: A simple and performant baseline for vision and language", arXiv preprint arXiv:1908.03557 (2019).

Li, X. L., A. Kuncoro, J. Hoffmann, C. de Masson d'Autume, P. Blunsom and A. Nematzadeh, "A systematic investigation of commonsense knowledge in large language models", in "Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing", pp. 11838–11855 (Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022b), URL https://aclanthology.org/2022.emnlp-main.812.

Liang, Y., N. Duan, Y. Gong, N. Wu, F. Guo, W. Qi, M. Gong, L. Shou, D. Jiang, G. Cao *et al.*, "Xglue: A new benchmark datasetfor cross-lingual pre-training, understanding and generation", in "Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)", pp. 6008–6018 (2020).

Lin, T.-Y., M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár and C. L. Zitnick, "Microsoft coco: Common objects in context", in "European conference on computer vision", pp. 740–755 (Springer, 2014).

Lin, X. and D. Parikh, "Leveraging visual question answering for image-caption ranking", in "Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14", pp. 261–277 (Springer, 2016).

Liu, D., Y. Yan, Y. Gong, W. Qi, H. Zhang, J. Jiao, W. Chen, J. Fu, L. Shou, M. Gong *et al.*, "Glge: A new general language generation evaluation benchmark", in "Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021", pp. 408–420 (2021).

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach", arXiv preprint arXiv:1907.11692 URL https://arxiv.org/pdf/1907.11692.pdf (2019).

Lobry, S., D. Marcos, J. Murray and D. Tuia, "Rsvqa: Visual question answering for remote sensing data", IEEE Transactions on Geoscience and Remote Sensing **58**, 12, 8555–8566 (2020).

Lu, J., D. Batra, D. Parikh and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks", Advances in neural information processing systems **32** (2019).

Lu, S., D. Guo, S. Ren, J. Huang, A. Svyatkovskiy, A. Blanco, C. Clement, D. Drain, D. Jiang, D. Tang *et al.*, "Codexglue: A machine learning benchmark dataset for code understanding and generation", in "Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)", (2021).

Luo, M., S. K. Sampat, R. Tallman, Y. Zeng, M. Vancha, A. Sajja and C. Baral, "'just because you are right, doesn't mean i am wrong': Overcoming a bottleneck in development and evaluation of open-ended vqa tasks", in "Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume", pp. 2766–2771 (2021).

Malinowski, M. and M. Fritz, "A multi-world approach to question answering about real-world scenes based on uncertain input", Advances in neural information processing systems **27** (2014).

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

McCarthy, J., "Situations, actions, and causal laws", Tech. rep., STANFORD UNIV CA DEPT OF COMPUTER SCIENCE (1963).

McCarthy, J. and P. Hayes, "Some philosophical problems from the standpoint of artificial intelligence", in "Machine Intelligence 4", edited by B. Meltzer and D. Michie, pp. 463–502 (Edinburgh University Press, 1969).

McCarthy, J. *et al.*, *Programs with common sense* (RLE and MIT computation center, 1960), URL `https://www.cs.rit.edu/~rlaz/is2014/files/McCarthyProgramsWithCommonSense.pdf`.

McHugh, M. L., "Interrater reliability: the kappa statistic", Biochemia medica **22**, 3, 276–282 (2012).

Mehri, S., M. Eric and D. Hakkani-Tur, "Dialoglue: A natural language understanding benchmark for task-oriented dialogue", arXiv preprint arXiv:2009.13570 (2020).

Mei, L., J. Mao, Z. Wang, C. Gan and J. B. Tenenbaum, "Falcon: Fast visual concept learning by integrating images, linguistic descriptions, and conceptual relations", in "International Conference on Learning Representations", (2021).

Menon, S. and C. Vondrick, "Visual classification via description from large language models", in "The Eleventh International Conference on Learning Representations", (2022).

Mishra, A., S. Shekhar, A. K. Singh and A. Chakraborty, "Ocr-vqa: Visual question answering by reading text in images", in "2019 international conference on document analysis and recognition (ICDAR)", pp. 947–952 (IEEE, 2019).

Mishra, S., A. Mitra, N. Varshney, B. Sachdeva, P. Clark, C. Baral and A. Kalyan, "Numglue: A suite of fundamental yet challenging mathematical reasoning tasks", in "Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)", pp. 3505–3523 (2022).

Mogadala, A., M. Kalimuthu and D. Klakow, "Trends in integration of vision and language research: A survey of tasks, datasets, and methods", Journal of Artificial Intelligence Research **71**, 1183–1317 (2021).

Moore, C., *The development of commonsense psychology* (Psychology Press, 2013).

Mukherjee, S., X. Liu, G. Zheng, S. Hosseini, H. Cheng, G. Yang, C. Meek, A. H. Awadallah and J. Gao, "Few-shot learning evaluation in natural language understanding", (2021).

Murphy, G. L., "What are categories and concepts", The making of human concepts pp. 11–28 (2010).

Murugesan, K., M. Atzeni, P. Kapanipathi, P. Shukla, S. Kumaravel, G. Tesauro, K. Talamadupula, M. Sachan and M. Campbell, "Text-based rl agents with commonsense knowledge: New challenges, environments and baselines", in "Proceedings of the AAAI Conference on Artificial Intelligence", vol. 35, pp. 9018–9027 (2021).

Naseem, U., M. Khushi and J. Kim, "Vision-language transformer for interpretable pathology visual question answering", IEEE Journal of Biomedical and Health Informatics **27**, 4, 1681–1690 (2023).

Nielsen, D., "Scandeval: A benchmark for scandinavian natural language processing", in "Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)", pp. 185–201 (2023).

OECD, "Pisa: Programme for international student assessment", Recuperado el URL `https://www.oecd-ilibrary.org/content/data/data-00365-en` (2019).

Otmazgin, S., A. Cattan and Y. Goldberg, "F-coref: Fast, accurate and easy to use coreference resolution", in "The 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (AACL-IJCNLP)", (2022).

Ott, M., S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling", in "Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)", pp. 48–53 (Association for Computational Linguistics, 2019), URL `https://www.aclweb.org/anthology/N19-4009`.

Papineni, K., S. Roukos, T. Ward and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation", in "Proceedings of the 40th annual meeting of the Association for Computational Linguistics", pp. 311–318 (2002).

Park, J., C. Bhagavatula, R. Mottaghi, A. Farhadi and Y. Choi, "Visualcomet: Reasoning about the dynamic context of a still image", in "Proceedings of the European Conference on Computer Vision (ECCV)", (2020), URL `https://arxiv.org/pdf/2004.10796.pdf`.

Park, S., J. Moon, S. Kim, W. I. Cho, J. Y. Han, J. Park, C. Song, J. Kim, Y. Song, T. Oh *et al.*, "Klue: Korean language understanding evaluation", in "Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)", (2021).

Pasupat, P. and P. Liang, "Compositional semantic parsing on semi-structured tables", in "Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)", (Association for Computational Linguistics, 2015).

Penedo, G., Q. Malartic, D. Hesslow, R. Cojocaru, A. Cappelli, H. Alobeidli, B. Pannier, E. Almazrouei and J. Launay, "The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only", arXiv preprint arXiv:2306.01116 (2023).

Peters, M. E., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee and L. Zettlemoyer, "Deep contextualized word representations", in "Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies", (2018).

Petroni, F., T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu and A. Miller, "Language models as knowledge bases?", in "Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)", pp. 2463–2473 (2019).

Pratt, S., R. Liu and A. Farhadi, "What does a platypus look like? generating customized prompts for zero-shot image classification", arXiv preprint arXiv:2209.03320 (2022).

Radford, A., J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision", in "International conference on machine learning", pp. 8748–8763 (PMLR, 2021).

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer", Journal of Machine Learning Research **21**, 140, 1–67, URL `http://jmlr.org/papers/v21/20-074.html` (2020).

Rajpurkar, P., R. Jia and P. Liang, "Know what you don't know: Unanswerable questions for squad", in "Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)", pp. 784–789 (2018).

Rajpurkar, P., J. Zhang, K. Lopyrev and P. Liang, "Squad: 100,000+ questions for machine comprehension of text", in "Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing", pp. 2383–2392 (2016).

Reddy, R. G., X. Rui, M. Li, X. Lin, H. Wen, J. Cho, L. Huang, M. Bansal, A. Sil, S.-F. Chang *et al.*, "Mumuqa: Multimedia multi-hop news question answering via cross-media knowledge extraction and grounding", in "Proceedings of the AAAI Conference on Artificial Intelligence", vol. 36, pp. 11200–11208 (2022).

Reinert, S., M. Hübener, T. Bonhoeffer and P. M. Goltstein, "Mouse prefrontal cortex represents learned rules for categorization", Nature **593**, 7859, 411–417 (2021).

Ren, M., R. Kiros and R. S. Zemel, "Exploring models and data for image question answering", in "Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada", pp. 2953–2961 (2015a), URL `https://proceedings.neurips.cc/paper/2015/hash/831c2f88a604a07ca94314b56a4921b8-Abstract.html`.

Ren, S., K. He, R. Girshick and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks", Advances in neural information processing systems **28** (2015b).

Roberts, A., C. Raffel and N. Shazeer, "How much knowledge can you pack into the parameters of a language model?", in "Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)", pp. 5418–5426 (2020).

Rybak, P., R. Mroczkowski, J. Tracz and I. Gawlik, "Klej: Comprehensive benchmark for polish language understanding", in "Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics", pp. 1191–1201 (2020).

Sampat, S., P. Banerjee, Y. Yang and C. Baral, "Learning action-effect dynamics for hypothetical vision-language reasoning task", Findings of EMNLP 2022. (2022a).

Sampat, S. K., A. Kumar, Y. Yang and C. Baral, "Clevr_hyp: A challenge dataset and baselines for visual question answering with hypothetical actions over images", in "Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies", pp. 3692–3709 (2021), URL `https://aclanthology.org/2021.naacl-main.289.pdf`.

Sampat, S. K., M. Nakamura, S. Kailas, K. Aggarwal, M. Zhou, Y. Yang and C. Baral, "Vl-glue: A suite of fundamental yet challenging visuo-linguistic reasoning tasks", arXiv preprint arXiv:2410.13666 (2024a).

Sampat, S. K., M. Patel, S. Das, Y. Yang and C. Baral, "Reasoning about actions over visual and linguistic modalities: A survey", arXiv preprint arXiv:2207.07568 (2022b).

Sampat, S. K., M. Patel, Y. Yang and C. Baral, "Help me identify: Is an llm+ vqa system all we need to identify visual concepts?", arXiv preprint arXiv:2410.13651 (2024b).

Sampat, S. K., Y. Yang and C. Baral, "Visuo-linguistic question answering (VLQA) challenge", in "Findings of the Association for Computational Linguistics: Empirical Methods in Natural Language Processing 2020", pp. 4606–4616 (Association for Computational Linguistics, 2020a), URL https://www.aclweb.org/anthology/2020.findings-emnlp.413.

Sampat, S. K., Y. Yang and C. Baral, "Visuo-lingustic question answering (vlqa) challenge", in "Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings", pp. 4606–4616 (2020b).

Sampat, S. K., Y. Yang and C. Baral, "Actioncomet: A zero-shot approach to learn image-specific commonsense concepts about actions", arXiv preprint arXiv:2410.13662 (2024c).

Sap, M., R. Le Bras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N. A. Smith and Y. Choi, "Atomic: An atlas of machine commonsense for if-then reasoning", in "Proceedings of the AAAI Conference on Artificial Intelligence", (2019).

Schick, T. and H. Schütze, "It's not just size that matters: Small language models are also few-shot learners", in "Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies", pp. 2339–2352 (2021).

Schwenk, D., A. Khandelwal, C. Clark, K. Marino and R. Mottaghi, "A-okvqa: A benchmark for visual question answering using world knowledge", in "European Conference on Computer Vision", pp. 146–162 (Springer, 2022).

Seelawi, H., I. Tuffaha, M. Gzawi, W. Farhan, B. Talafha, R. Badawi, Z. Sober, O. Al-Dweik, A. A. Freihat and H. Al-Natsheh, "Alue: Arabic language understanding evaluation", in "Proceedings of the Sixth Arabic Natural Language Processing Workshop", pp. 173–184 (2021).

Seo, M., H. Hajishirzi, A. Farhadi and O. Etzioni, "Diagram understanding in geometry questions", in "AAAI Conference on Artificial Intelligence", (2014).

Seo, M., A. Kembhavi, A. Farhadi and H. Hajishirzi, "Bidirectional attention flow for machine comprehension", in "International Conference on Learning Representations", (2016).

Shah, S., A. Mishra, N. Yadati and P. P. Talukdar, "Kvqa: Knowledge-aware visual question answering", in "Proceedings of the AAAI Conference on Artificial Intelligence", vol. 33, pp. 8876–8884 (2019).

Shannon, K., "The word on water: Conserve", The Dallas Morning News URL https://in.pinterest.com/pin/64035625923787308/ (2013).

Shavrina, T., A. Fenogenova, E. Anton, D. Shevelev, E. Artemova, V. Malykh, V. Mikhailov, M. Tikhonova, A. Chertok and A. Evlampiev, "Russiansuperglue: A russian language understanding evaluation benchmark", in "Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)", pp. 4717–4726 (2020).

Shen, T., M. Ott, M. Auli and M. Ranzato, "Mixture models for diverse machine translation: Tricks of the trade", in "Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA", vol. 97 of *Proceedings of Machine Learning Research*, pp. 5719–5728 (PMLR, 2019), URL http://proceedings.mlr.press/v97/shen19c.html.

Shon, S., A. Pasad, F. Wu, P. Brusco, Y. Artzi, K. Livescu and K. J. Han, "Slue: New benchmark tasks for spoken language understanding evaluation on natural speech", in "ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)", pp. 7927–7931 (IEEE, 2022).

Singh, A., V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh and M. Rohrbach, "Towards vqa models that can read", in "Proceedings of the IEEE/CVF conference on computer vision and pattern recognition", pp. 8317–8326 (2019).

Singh, H., A. Nasery, D. Mehta, A. Agarwal, J. Lamba and B. V. Srinivasan, "Mimoqa: Multimodal input multimodal output question answering", in "Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies", pp. 5317–5332 (2021).

Speer, R., J. Chin and C. Havasi, "Conceptnet 5.5: An open multilingual graph of general knowledge", in "Proceedings of the AAAI conference on artificial intelligence", vol. 31 (2017).

Storks, S., Q. Gao, Y. Zhang and J. Chai, "Tiered reasoning for intuitive physics: Toward verifiable commonsense language understanding", in "In Findings of the Conference on Empirical Methods in Natural Language Processing", (2021).

Su, L., N. Duan, E. Cui, L. Ji, C. Wu, H. Luo, Y. Liu, M. Zhong, T. Bharti and A. Sacheti, "Gem: A general evaluation benchmark for multimodal tasks", in "Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021", pp. 2594–2603 (2021).

Su, W., X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei and J. Dai, "Vl-bert: Pre-training of generic visual-linguistic representations", in "International Conference on Learning Representations", (2019).

Suhr, A., M. Lewis, J. Yeh and Y. Artzi, "A corpus of natural language for visual reasoning", in "Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Vol 2: Short Papers)", (2017).

Suhr, A., S. Zhou, A. Zhang, I. Zhang, H. Bai and Y. Artzi, "A corpus for reasoning about natural language grounded in photographs", in "Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics", pp. 6418–6428 (2019).

Talmor, A., O. Yoran, A. Catav, D. Lahav, Y. Wang, A. Asai, G. Ilharco, H. Hajishirzi and J. Berant, "Multimodalqa: complex question answering over text, tables and images", in "International Conference on Learning Representations", (2020).

Tan, H. and M. Bansal, "LXMERT: Learning cross-modality encoder representations from transformers", in "Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)", pp. 5100–5111 (Association for Computational Linguistics, 2019), URL `https://www.aclweb.org/anthology/D19-1514`.

Tandon, N., B. Dalvi, K. Sakaguchi, P. Clark and A. Bosselut, "WIQA: A dataset for "what if..." reasoning over procedural text", in "Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)", pp. 6076–6085 (Association for Computational Linguistics, 2019), URL `https://www.aclweb.org/anthology/D19-1629`.

Tandon, N., G. Melo and G. Weikum, "Acquiring comparative commonsense knowledge from the web", in "Proceedings of the AAAI Conference on Artificial Intelligence", vol. 28 (2014).

Thrush, T., R. Jiang, M. Bartolo, A. Singh, A. Williams, D. Kiela and C. Ross, "Winoground: Probing vision and language models for visio-linguistic compositionality", in "Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition", pp. 5238–5248 (2022).

Touvron, H., T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models", arXiv preprint arXiv:2302.13971 (2023).

Trott, A., C. Xiong and R. Socher, "Interpretable counting for visual question answering", in "International Conference on Learning Representations", (2018).

Valmeekam, K., A. Olmo, S. Sreedharan and S. Kambhampati, "Large language models still can't plan (a benchmark for llms on planning and reasoning about change)", in "Advances in Neural Information Processing Systems 2022 Foundation Models for Decision Making Workshop", (2022).

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, "Attention is all you need", Advances in neural information processing systems **30** (2017).

Vedantam, R., C. Lawrence Zitnick and D. Parikh, "Cider: Consensus-based image description evaluation", in "Proceedings of the IEEE conference on computer vision and pattern recognition", pp. 4566–4575 (2015).

Vo, N., L. Jiang, C. Sun, K. Murphy, L. Li, L. Fei-Fei and J. Hays, "Composing text and image for image retrieval - an empirical odyssey", in "IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019", pp. 6439–6448 (Computer Vision Foundation / IEEE, 2019), URL http://openaccess.thecvf.com/content_CVPR_2019/html/Vo_Composing_Text_and_Image_for_Image_Retrieval_-_an_Empirical_CVPR_2019_paper.html.

Vrandečić, D. and M. Krötzsch, "Wikidata: a free collaborative knowledgebase", Communications of the ACM **57**, 10, 78–85 (2014).

Wagner, M., H. Basevi, R. Shetty, W. Li, M. Malinowski, M. Fritz and A. Leonardis, "Answering visual what-if questions: From actions to predicted scene descriptions", in "Proceedings of the European Conference on Computer Vision (ECCV)", pp. 0–0 (2018), URL https://openaccess.thecvf.com/content_ECCVW_2018/papers/11129/Wagner_Answering_Visual_em_What-If_Questions_From_Actions_to_Predicted_Scene_ECCVW_2018_paper.pdf.

Wah, C., S. Branson, P. Welinder, P. Perona and S. Belongie, "The caltech-ucsd birds-200-2011 dataset", Computation & Neural Systems Technical Report (2011).

Wang, A., Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy and S. Bowman, "Superglue: A stickier benchmark for general-purpose language understanding systems", Advances in neural information processing systems **32** (2019a).

Wang, A., A. Singh, J. Michael, F. Hill, O. Levy and S. Bowman, "Glue: A multi-task benchmark and analysis platform for natural language understanding", in "Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP", pp. 353–355 (2018).

Wang, A., A. Singh, J. Michael, F. Hill, O. Levy and S. R. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding", in "International Conference on Learning Representations", (2019b), URL https://openreview.net/forum?id=rJ4km2R5t7.

Wang, B., C. Xu, S. Wang, Z. Gan, Y. Cheng, J. Gao, A. H. Awadallah and B. Li, "Adversarial glue: A multi-task benchmark for robustness evaluation of language models", in "Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)", (2021).

Wang, J., Z. Yang, X. Hu, L. Li, K. Lin, Z. Gan, Z. Liu, C. Liu and L. Wang, "Git: A generative image-to-text transformer for vision and language", Transactions on Machine Learning Research (2022a).

Wang, P., Q. Wu, C. Shen, A. Dick and A. Van Den Henge, "Explicit knowledge-based reasoning for visual question answering", in "Proceedings of the 26th International Joint Conference on Artificial Intelligence", pp. 1290–1296 (2017a).

Wang, P., Q. Wu, C. Shen, A. Dick and A. Van Den Hengel, "Fvqa: Fact-based visual question answering", IEEE transactions on pattern analysis and machine intelligence **40**, 10, 2413–2427 (2017b).

Wang, Y., S. Mishra, P. Alipoormolabashi, Y. Kordi, A. Mirzaei, A. Arunkumar, A. Ashok, A. S. Dhanasekaran, A. Naik, D. Stap *et al.*, "Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks", arXiv preprint arXiv:2204.07705 (2022b).

Wu, J., Z. Hu and R. Mooney, "Generating question relevant captions to aid visual question answering", in "Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics", pp. 3585–3594 (2019).

Xie, N., F. Lai, D. Doran and A. Kadav, "Visual entailment: A novel task for fine-grained image understanding", arXiv preprint arXiv:1901.06706 (2019).

Xu, L., H. Hu, X. Zhang, L. Li, C. Cao, Y. Li, Y. Xu, K. Sun, D. Yu, C. Yu *et al.*, "Clue: A chinese language understanding evaluation benchmark", in "Proceedings of the 28th International Conference on Computational Linguistics", pp. 4762–4772 (2020).

Xu, L., X. Lu, C. Yuan, X. Zhang, H. Xu, H. Yuan, G. Wei, X. Pan, X. Tian, L. Qin *et al.*, "Fewclue: A chinese few-shot learning evaluation benchmark", arXiv preprint arXiv:2107.07498 (2021).

Yagcioglu, S., A. Erdem, E. Erdem and N. Ikizler-Cinbis, "Recipeqa: A challenge dataset for multimodal comprehension of cooking recipes", in "Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing", pp. 1358–1368 (2018).

Yan, A., Y. Wang, Y. Zhong, C. Dong, Z. He, Y. Lu, W. Wang, J. Shang and J. McAuley, "Learning concise and descriptive attributes for visual recognition", arXiv preprint arXiv:2308.03685 (2023).

Yang, G. R., I. Ganichev, X.-J. Wang, J. Shlens and D. Sussillo, "A dataset and architecture for visual reasoning with a working memory", in "Proceedings of the European Conference on Computer Vision (ECCV)", pp. 714–731 (2018), URL `https://www.ecva.net/papers/eccv_2018/papers_ECCV/papers/Guangyu_Robert_Yang_A_dataset_and_ECCV_2018_paper.pdf`.

Yang, L., S. Zhang, L. Qin, Y. Li, Y. Wang, H. Liu, J. Wang, X. Xie and Y. Zhang, "GLUE-X: Evaluating natural language understanding models from an out-of-distribution generalization perspective", in "Findings of the Association for Computational Linguistics: ACL 2023", pp. 12731–12750 (Association for Computational Linguistics, Toronto, Canada, 2023a), URL `https://aclanthology.org/2023.findings-acl.806`.

Yang, Y., A. Panagopoulou, Q. Lyu, L. Zhang, M. Yatskar and C. Callison-Burch, "Visual goal-step inference using wikiHow", in "Proceedings of the 2021 Conference

on Empirical Methods in Natural Language Processing", pp. 2167–2179 (Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021a), URL https://aclanthology.org/2021.emnlp-main.165.

Yang, Y., A. Panagopoulou, Q. Lyu, L. Zhang, M. Yatskar and C. Callison-Burch, "Visual goal-step inference using wikihow", in "Proceedings of the Conference on Empirical Methods in Natural Language Processing", (2021b).

Yang, Y., A. Panagopoulou, S. Zhou, D. Jin, C. Callison-Burch and M. Yatskar, "Language in a bottle: Language model guided concept bottlenecks for interpretable image classification", in "Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition", pp. 19187–19197 (2023b).

Yang, Z., X. He, J. Gao, L. Deng and A. Smola, "Stacked attention networks for image question answering", in "Proceedings of the IEEE conference on computer vision and pattern recognition", pp. 21–29 (2016).

Yeo, J., G. Lee, G. Wang, S. Choi, H. Cho, R. K. Amplayo and S.-w. Hwang, "Visual choice of plausible alternatives: An evaluation of image-based commonsense causal reasoning", in "Proceedings of the International Conference on Language Resources and Evaluation", (2018).

Yi, K., J. Wu, C. Gan, A. Torralba, P. Kohli and J. Tenenbaum, "Neural-symbolic VQA: disentangling reasoning from vision and language understanding", in "Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada", pp. 1039–1050 (2018), URL https://proceedings.neurips.cc/paper/2018/hash/5e388103a391daabe3de1d76a6739ccd-Abstract.html.

Yin, Z., H. Wang, K. Horio, D. Kawahara and S. Sekine, "Should we respect llms? a cross-lingual study on the influence of prompt politeness on llm performance", arXiv preprint arXiv:2402.14531 (2024).

Zellers, R., Y. Bisk, A. Farhadi and Y. Choi, "From recognition to cognition: Visual commonsense reasoning", in "Proceedings of the IEEE/CVF conference on computer vision and pattern recognition", pp. 6720–6731 (2019).

Zhang, N., M. Chen, Z. Bi, X. Liang, L. Li, X. Shang, K. Yin, C. Tan, J. Xu, F. Huang *et al.*, "Cblue: A chinese biomedical language understanding evaluation benchmark", in "Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)", pp. 7888–7915 (2022).

Zhou, L., N. Louis and J. J. Corso, "Weakly-supervised video object grounding from text by loss weighting and object interaction", in "British Machine Vision Conference", (2018a), URL http://bmvc2018.org/contents/papers/0070.pdf.

Zhou, L., C. Xu and J. J. Corso, "Towards automatic learning of procedures from web instructional videos", in "AAAI Conference on Artificial Intelligence", pp. 7590–7598 (2018b), URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17344.

Zhu, Y., O. Groth, M. S. Bernstein and L. Fei-Fei, "Visual7w: Grounded question answering in images", in "2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016", pp. 4995–5004 (IEEE Computer Society, 2016), URL https://doi.org/10.1109/CVPR.2016.540.