A Appendices

A.1 Survey of existing vision-Language Datasets and Comparison with VLQA

There are several datasets proposed in recent years to benchmark a variety of vision-language tasks. We provide the list of such datasets and comparison with our dataset in Table 5 based on following attributes;

1. **Dataset** name with corresponding url of dataset website/publication is available.

2. **Modality** states which of the following components each dataset has; I (Images), T (Text as a QA mechanism), T+ (Additional Textual Context), K (Additional Knowledge required). VLQA dataset has all four components, standing out from the rest of the datasets except TQA, which has a different objective of textbook-style learning. This makes VLQA task harder than other existing datasets, which we believe will be a driver for development of more advanced AI models.

3. **Visual Modality Classification** describes the nature of visuals incorporated for a dataset which are categorized in 3 major kinds; Natural (everyday objects and scenes), Synthetic (artificially/program generated or templated figures) or Diagrams (imagery representing complex relationships between multiple interrelated objects or phenomena). Our dataset includes all three kinds of visuals aiming at developing generic vision-language reasoning system. However, we provide this classification as a part of our annotations for researchers interested in advancements specific to a particular kind of visual.

4. **Textual Modality Classification** describes the nature of language component incorporated for a dataset. Most commonly used texts are in the form of Question, Caption, Sentence with exceptions of a Lesson and a Paragraph in TQA and VLQA respectively.

5. **Task** represents the broad categorization defined by the NLP and Computer Vision community for each vision-language problem. Most tasks are in the form of question answering, popularly known as VQA. Additionally, if the task focuses on a particular reasoning skill needed to solve the dataset (e.g. counting, spatial reasoning, understanding text within images) or requires a domain specific knowledge, (charts, science, geometry, commonsense or world knowledge) is mentioned alongside.

6. **Task Types** indicates whether a task can be solved as a Classification, Text generation or a Ranking problem. Classification tasks are commonly formed as a Multiple Choice (MC) or N-class classification. Open Ended (OE) answers (as strings or numeric) and Captions are standard mechanisms to evaluate text generation style tasks. Vision-language task for ShapeWorld is the only one which employs Scoring mechanism to represent confidence level in range [0,1].

A.2 Dataset Creation Pipeline

Figure 6 illustrates the complete dataset creation pipeline. We divide overall process in 3 main stages- Data Collection, Annotation and Quality Control which is explained below;

A.2.1 Data Collection and Post-processing

VLQA task requires <Image, Passage, Question, AnswerChoices> for each item in the corpus. To curate this dataset, we rely on data collection in two ways; One where variety of images are collected through crawling scripts that uses keyword search, existing APIs (flickr, twitter, newspapers, wikipedia, infographic websites etc.), images collected from documents and encyclopedias, which we refer to as primary data source. Then we manually find the relevant textual information in the context of the image and create questions based on it. We also tried generating templated images (like bar chart, pie chart, scatter plot etc.) from the tabular data obtained from CIA 'world factbook' and WikiTables dataset. In the secondary data based method, we directly import items from human psychometric tests, exercises from school textbooks/handouts or existing vision-language datasets and then modify it in a way so that it fits the VLQA task. The data collection process included writing crawling/scraping scripts followed by combination of manual and automated search and fix such as,

- replacing given textual/visual data with equivalent visual/textual counterparts respectively
- adding/removing partial information to/from text or visuals so that image and text do not contain identical information

Dataset	Modality				Modality Classification		Task Type	Task (Domain)
	I	T	T+	K	Visual	Textual		
Clevr	✓	✓	✗	✗	Synthetic	Ques	OE	VQA (Spatial Reasoning)
COCO	✓	✓	✗	✗	Natural	Caption	Caption	Text generation
COCO-BISON	✓	✓	✗	✗	Natural	Sent	MC	Image Selection
COCO-QA	✓	✓	✗	✗	Natural	Ques	OE	VQA
COG	✓	✓	✗	✗	Synthetic	Ques / Sent	MC	VQA, Instruction Following
Concept.Caption	✓	✓	✗	✗	Natural	Caption	Caption	Text generation
CountQA	✓	✓	✗	✗	Natural	Ques	Numeral	VQA (Counting)
DAQUAR	✓	✓	✗	✗	Natural	Ques	OE	VQA
DVQA	✓	✓	✗	✗	Synthetic	Ques	OE	VQA (BarCharts)
FigureQA	✓	✓	✗	✗	Synthetic	Ques	OE	VQA (Charts)
FMIQA	✓	✓	✗	✗	Natural	Ques	OE	VQA
GQA	✓	✓	✗	✗	Natural	Ques	OE	VQA
HowManyQA	✓	✓	✗	✗	Natural	Ques	Numeral	VQA (Counting)
LEAFQA	✓	✓	✗	✗	Synthetic	Ques	OE	VQA (Charts)
Memex-QA	✓	✓	✗	✗	Natural	Ques	MC	VQA
MSRVTT-QA	✓	✓	✗	✗	Natural	Ques	OE	VQA
NLVRv1/v2	✓	✓	✗	✗	Synthetic/Natural	Sent	T/F	Text classification
OpenImagesV6	✓	✓	✗	✗	Natural	Caption	Caption	Text generation
RVQA	✓	✓	✗	✗	Natural	Ques	OE	VQA
Shapes	✓	✓	✗	✗	Synthetic	Ques	OE	VQA
ShapeWorld	✓	✓	✗	✗	Synthetic	Sent	Scoring	Text classification
SNLI-VE	✓	✓	✗	✗	Natural	Sent	3 classes	Visual Entailment
TallyQA	✓	✓	✗	✗	Natural	Ques	Numeric	VQA (Counting)
TDIUC	✓	✓	✗	✗	Natural	Ques	OE	VQA
TextVQA	✓	✓	✗	✗	Natural	Ques	OE	VQA (Text in Images)
VCR	✓	✓	✗	✗	Natural	Ques	MC	VQA+Rationale
Vis.Genome	✓	✓	✗	✗	Natural	Ques	OE	VQA (Scene Graphs)
Vis.Madlibs	✓	✓	✗	✗	Natural	Sent	Blanks	VQA
Vis.7W	✓	✓	✗	✗	Natural	Ques	MC	VQA
Vis.Dialogue	✓	✓	✗	✗	Natural	Ques	OE	VQA (Dialogue)
VizWiz-Priv	✓	✓	✗	✗	Natural	Ques	OE	VQA (Text in Images)
VQAv1 Abs./Real	✓	✓	✗	✗	Synthetic/Natural	Ques	OE	VQA
VQAv2/CP	✓	✓	✗	✗	Natural	Ques	OE,MC	VQA
WAT2019	✓	✓	✗	✗	Natural	Caption	Caption	Text generation / Translation
AI2 Geometry	✓	✓	✗	✓	Diagrams	Ques	MC	VQA (Geometry)
AI2 Mercury	✓	✓	✗	✓	Diagrams	Ques	MC	VQA (Science)
AI2 ScienceQ	✓	✓	✗	✓	Diagrams	Ques	MC	VQA (Science)
AI2D	✓	✓	✗	✓	Diagrams	Ques	MC	VQA (Science)
FVQA	✓	✓	✗	✓	Natural	Ques	OE	VQA (Commonsense)
KBVQA	✓	✓	✗	✓	Natural	Ques	OE	VQA (Commonsense)
KVQA	✓	✓	✗	✓	Natural	Ques	OE	VQA (World Knowledge)
OKVQA	✓	✓	✗	✓	Natural	Ques	OE	VQA (World Knowledge)
WKVQA	✓	✓	✗	✓	Natural	Ques	OE	VQA (World Knowledge)
TQA	✓	✓	✓	✓	Diagrams	Ques, Lesson	MC	VQA (Science)
VLQA (Our)	✓	✓	✓	✓	Natural, Synthetic, Diagrams	Ques, Para	MC	VQA (Joint Reasoning over Image-Text)

Table 5: Survey of existing vision-Language Datasets and Comparison with VLQA

- creating factual or hypothetical situations around images

Then we standardize all collected information using above methods as multiple choice question-answers (MCQs) and get the initial version of the dataset. Our dataset includes all three kinds of visuals- Natural (everyday objects and scenes), Synthetic (artificially/program generated or templated figures) or Diagrams (imagery representing complex relationships between objects or phenomena). Each item in the VLQA dataset involves a considerable amount of text in passage, question and answer choices formed of diverse vocabulary of 33259 unique tokens. Also, these texts can involve facts, imaginary scenarios or their combination making it more realistic for real-world scenarios. This is how we compiled a large number of diverse items for the VLQA dataset in order to develop a generic vision-language reasoning system.

A.2.2 Annotation

All data items obtained from primary sources require annotation as questions are created manually. For items obtained through secondary methods, annotation is required only if originally imported modalities were perturbed. Our crowd worker interface was designed using Python-Flask[5] and deployed on a local server. The data entered by annotators is then logged into Comma Separated Value (CSV) files in a structured format. Annotators were clearly instructed (as per figure 7) about the annotation procedure and it was known to them that exactly one answer choice is correct for each item in our dataset. During first round of annotations, annotators were allowed to reject bad samples based on following two things; first, image and passage must not represent identical information and second, a question must not be answerable without looking at image and passage by marking it ambiguous as shown in Figure 8.

A.2.3 Quality Control and Bias Mitigation

Since we focus on the task of joint reasoning, we have to ensure that all our data items must use both image and passage. For the quality control purposes, we want to remove the samples which can be answered correctly by the state-of-the-art models in the absence of one of the modalities due to underlying bias it has learned from the training

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

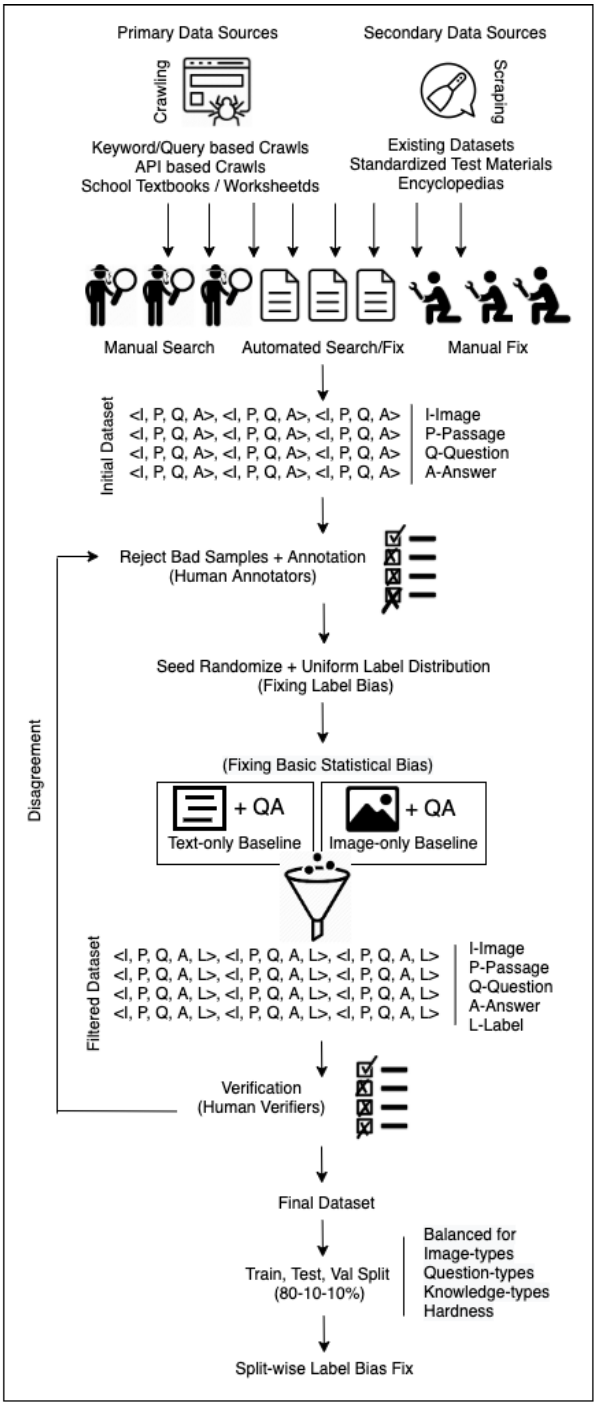
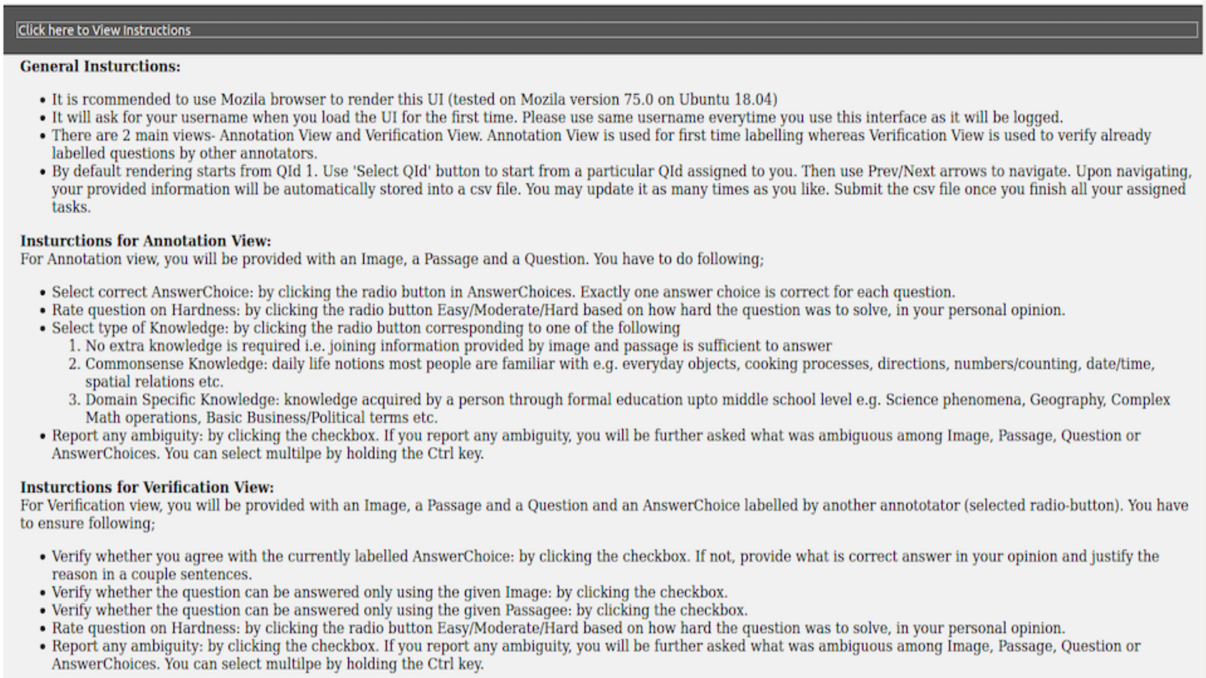


Figure 6: Data Collection, Processing and Integrity Steps implemented for construction of VLQA

data. Therefore, we create 3 baselines- question-only (simply takes Q and predicts answer from choices A), passage-only (considers P as a context, takes Q and predicts answer from choices A) and image-only (considers I as a context, takes Q and predicts answer from choices A). We get predictions for whole data using these baselines. We repeat this experiment for 3 times by shuffling answer choices with a fixed seed. If a question can



Click here to View Instructions

General Insturctions:

- It is rcommended to use Mozila browser to render this UI (tested on Mozila version 75.0 on Ubuntu 18.04)
- It will ask for your username when you load the UI for the first time. Please use same username everytime you use this interface as it will be logged.
- There are 2 main views- Annotation View and Verification View. Annotation View is used for first time labelling whereas Verification View is used to verify already labelled questions by other annotators.
- By default rendering starts from QId 1. Use 'Select QId' button to start from a particular QId assigned to you. Then use Prev/Next arrows to navigate. Upon navigating, your provided information will be automatically stored into a csv file. You may update it as many times as you like. Submit the csv file once you finish all your assigned tasks.

Insturctions for Annotation View:
For Annotation view, you will be provided with an Image, a Passage and a Question. You have to do following;

- Select correct AnswerChoice: by clicking the radio button in AnswerChoices. Exactly one answer choice is correct for each question.
- Rate question on Hardness: by clicking the radio button Easy/Moderate/Hard based on how hard the question was to solve, in your personal opinion.
- Select type of Knowledge: by clicking the radio button corresponding to one of the following
  1. No extra knowledge is required i.e. joining information provided by image and passage is sufficient to answer
  2. Commonsense Knowledge: daily life notions most people are familiar with e.g. everyday objects, cooking processes, directions, numbers/counting, date/time, spatial relations etc.
  3. Domain Specific Knowledge: knowledge acquired by a person through formal education upto middle school level e.g. Science phenomena, Geography, Complex Math operations, Basic Business/Political terms etc.
- Report any ambiguity: by clicking the checkbox. If you report any ambiguity, you will be further asked what was ambiguous among Image, Passage, Question or AnswerChoices. You can select multilpe by holding the Ctrl key.

Insturctions for Verification View:
For Verification view, you will be provided with an Image, a Passage and a Question and an AnswerChoice labelled by another annototator (selected radio-button). You have to ensure following;

- Verify whether you agree with the currently labelled AnswerChoice: by clicking the checkbox. If not, provide what is correct answer in your opinion and justify the reason in a couple sentences.
- Verify whether the question can be answered only using the given Image: by clicking the checkbox.
- Verify whether the question can be answered only using the given Passagee: by clicking the checkbox.
- Rate question on Hardness: by clicking the radio button Easy/Moderate/Hard based on how hard the question was to solve, in your personal opinion.
- Report any ambiguity: by clicking the checkbox. If you report any ambiguity, you will be further asked what was ambiguous among Image, Passage, Question or AnswerChoices. You can select multilpe by holding the Ctrl key.

Figure 7: **3-fold instructions for annotators** Generic Instructions, Annotation Instructions and Verification Instructions.

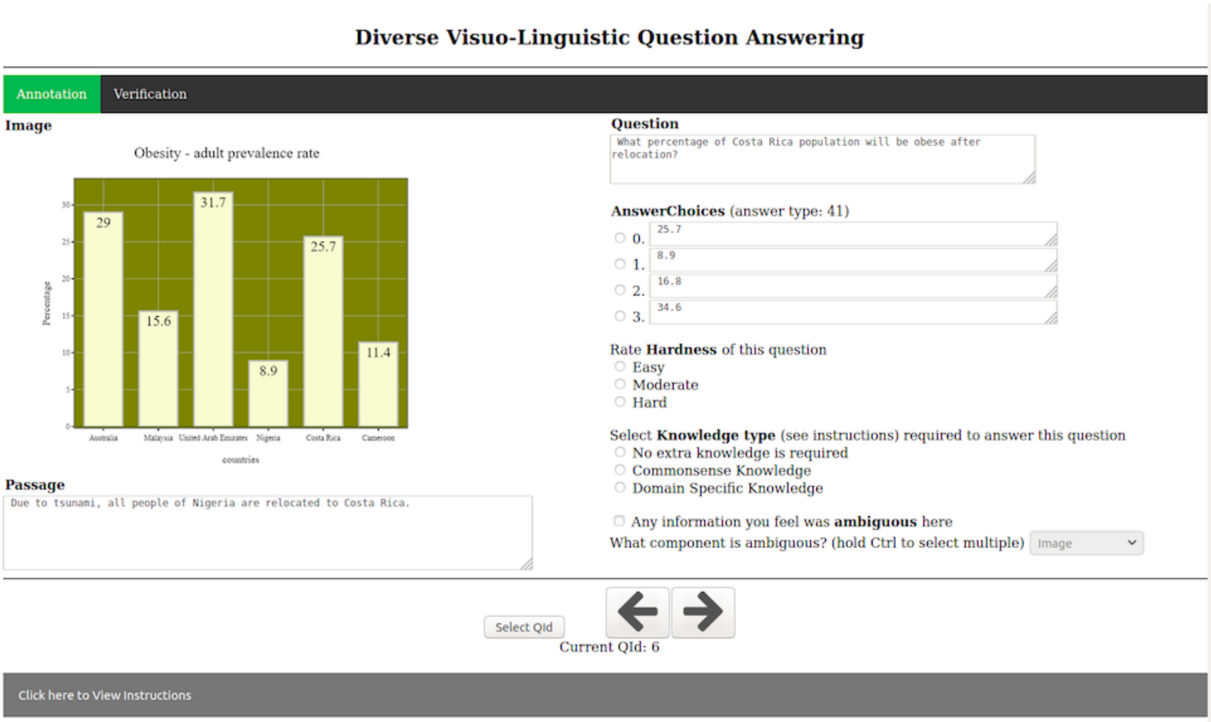


Figure 8: **Annotation View** is used for first time labelling of dataset items. <I, P, Q, A> will be rendered in the UI after initial dataset formation. User has to determine the correct choice, categorize item based on knowledge type, rate for hardness and report ambiguity (if any).

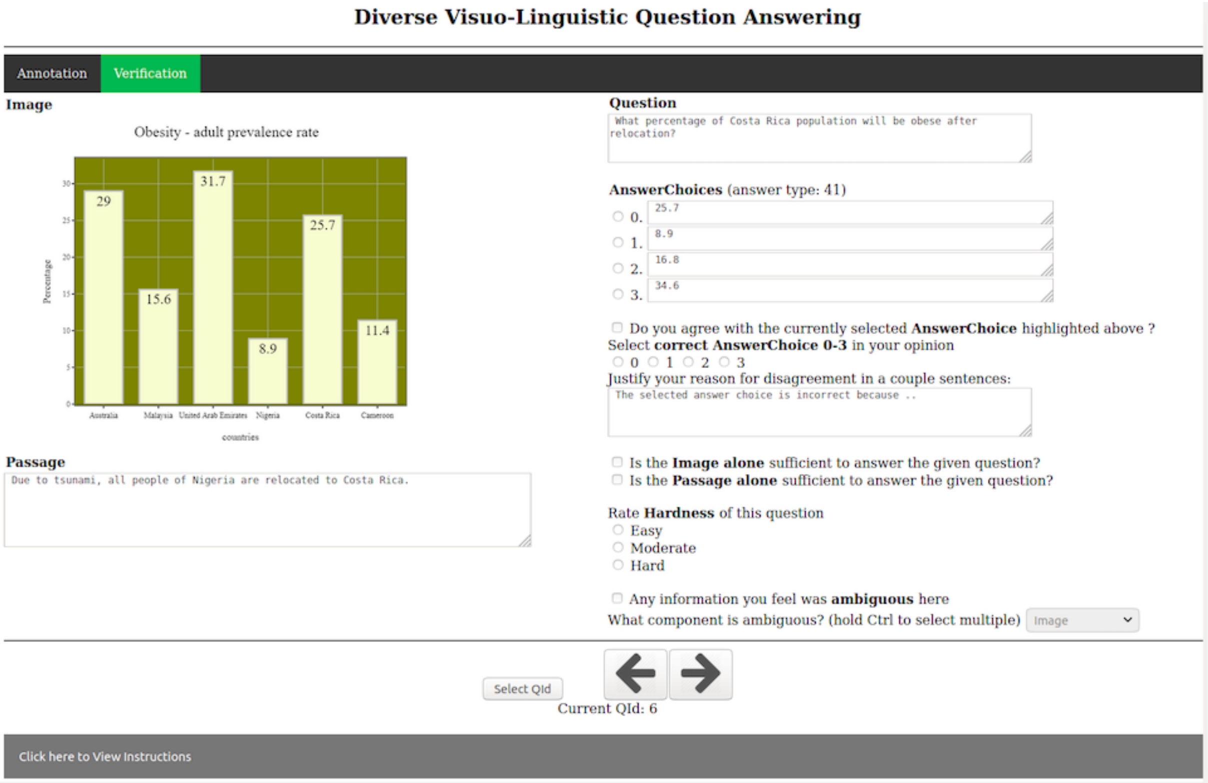


Figure 9: **Verification View** is used as a mechanism for inter-annotator agreement about the ground-truth label. <I, P, Q, A, L> will be rendered in the UI post image-only and text-only baseline filtering. User checks for the correctness of label, rate for hardness and report ambiguity (if any).

be answered correctly by any baseline in all trials, We remove such samples. Performance for these baselines is reported in Table 3. The poor performance of these baselines indicate that the VLQA dataset requires models to jointly understand both image and text modalities.

Finally, we perform another round of manual quality check. We instruct workers to first try to answer a question just based on images and then try to answer a question based on only using text passage. If a question can be answered using a single modality, we suggest annotators to mark the checkbox as shown in Figure 9. Finally, we look over all bad samples and either provide a fix or remove, on a case-by-case basis.

We initially curated ~12000 image-passage-qa pairs. During the annotation process, ~700 were reported ambiguous, out of which we removed ~500 and remaining ~200 were modified and added back. By quality check process through baselines, we removed another ~1900 samples. In the verification stage, we further removed ~350 samples, and ended up with a dataset of 9267 samples eventually. Two rejected examples can be seen in Figure 10, with explanation of reason for removal.

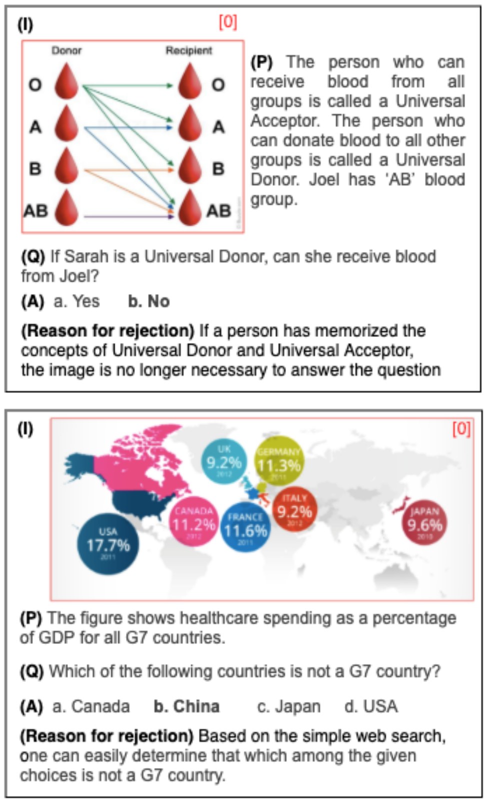


Figure 10: Example of 2 rejected VLQA samples with explanation for rejection.

A.3 Format of Annotations provided for VLQA Dataset and explanation of each field

```
{
  "qid": 1,
  "images": [1.png,2.png,..],
  "multiple_images" : True/False,
  "passage": "This is a sample text passage.",
  "question": "Is this a sample question?",
  "answer_choices": ["choice0", "choice1", "choice2", "choice3"],
  "answer": 0/1/2/3,
  "image_type": "Natural"/"Templated"/"Freeform"
  "image_subtype": "Bar"/"Pie"/..,
  "answer_type": "4way_text",
  "multistep_inference": True/False,
  "reasoning_type": ["Deductive","Math"],
  "ext_knowledge": True/False,
  "ext_knowledge_type": "Commonsense"
  "ext_knowledge_text": "This is external knowledge required.",
  "ocrtokens": ["text","tokens","inside","image"],
  "image_source": "http://www.image/obtained/from/url/xyz",
  "passage_source": "wikipedia",
  "difficulty_level": "hard"/"easy"/"moderate",
  "split": "train"/"test"/"val"
}
```

- **qid**: Unique identifier for the item from 1 to 9267
- **images**: Visual modality for the dataset item as a list of image file names, which will be assigned unique identifiers [0],[1],[2],.. and composed as a single file by merging (in order left to right)
- **multiple_images**: Boolean field suggesting whether or not an item has multiple images
- **passage**: Textual modality for the dataset item, typically consisting of 1-5 sentences.
- **question**: Question in natural language aiming to assess joint reasoning capability of a person/model
- **answer_choices**: Answer choices for a multiple choice question (MCQ) which can be short phrases, numeric, sentence, boolean or image (referred as a detection tag [0],[1],[2],..)
- **answer**: Integer 0-3 corresponding to answer_choices suggesting the ground-truth label for a question
- **image_type**: Categorization of images based on whether they are "Natural", "Templated" (structured) or "Freeform" (unstructured and not natural)
- **image_subtype**: "Templated" images are further classified in 20 subtypes listed as follows; "Bar" (includes Simple/Stacked/Grouped), "Pie" (or Donut chart), "Scatter", "Line", "Area", "Bubble" , "Radar", "VennDiagrams", "Timelines", "Hierarchies" (or Trees), "Maps", "Tables" (or Matrix), "Cycles", "Processes", "Heatmaps", "DirectedGraphs", "UndirectedGraphs", "FlowCharts", "SankeyDiagram", "CoordinateSystems" (this field will be empty for "Natural" and "Freeform" images)
- **answer_type**: Classification of item based on 5 answer types listed as follows;
  1. **4-way text (4wT)**: ["text0","text1","text2","text3"]
     One need to select the correct alpha-numeric choice among 4 choices based on the scenario described in question, passage and image

2. **4-way Sequencing (4wS)**: [“I-II-IV-III”,“I-IV-III-II”,“II-III-I-IV”,“II-I-IV-III”]
   Consider 4 steps (I-IV) in a process which is represented as a combination of image and text, and jumbled up. One has to select the correct order of events from given choices.
3. **4-way image (4wI)**: [“[1]”,“[2]”,“[0]”,“[3]”] where [x] are image detection tags
   One needs to select the correct image among 4 choices based on the scenario described in question and passage. Images are referred through detection tags [0],[1],[2],[3].
4. **2-way image (2wI)**: [“[1]”,“[0]”] where [x] are image detection tags
   One needs to select the correct image among 2 choices based on the scenario described in question and passage. Images are referred through detection tags [0],[1].
5. **Binary Classification (Bin)**: [“True”,“False”] or [“No”,“Yes”]
   One needs to determine whether or not the text in question is true or false with respect to the given visuo-linguistic context.

- **multistep_inference**: Boolean field suggesting whether or not the question requires multiple inference steps to correctly answer the question

- **reasoning_type**: A list of reasoning skills required to solve given question, most frequently observed types are listed as follows;

  1. “InfoLookup” (look for a specific information or conditional retrieval)
  2. “Temporal” (reasoning with respect to time)
  3. “Spatial” (reasoning with respect to space)
  4. “Deductive” (given a generic principle, deduce a conclusion for a specific case and vice-versa)
  5. “Abductive” (finding most plausible explanation with respect to given set of observations)
  6. “Mathematical” (arithmetic, trends, minimum/maximum, counting, comparison, complement, fractions, percentages etc.)
  7. “Logical” (conjunction, disjunction, logical negation, existential quantifiers etc.)
  8. “Causality” (cause-effect relationship)
  9. “Analogy” (comparison for the purpose of explanation or clarification, different from numerical comparison)
  10. “Verbal” (synonym/antonym, subclass/superclass, vocabulary, verbal negation etc.)

- **ext_knowledge**: Boolean field suggesting whether or not the question requires any external knowledge beyond what is provided in visuo-linguistic context

- **ext_knowledge_type**: Classification of required external knowledge as follows;

  1. Commonsense: Facts about the everyday world, which most people know.
  2. Domain Specific knowledge: Knowledge acquired through formal study (we limit our domain specific knowledge to middle-school level)

- **ext_knowledge_text**: Manually written justification of required external knowledge

- **ocrtokens**: List of OCR extracted tokens from images and manually corrected if erroneous, just in case some systems would like incorporate OCR based features

- **image_source**: Source/Weblink from which image is retrieved (original image might be altered in some cases before using it for this dataset)

- **passage_source**: Source/Weblink of passage (if retrieved from some source), empty if passage written manually

- **difficulty_level**: Difficulty level of question, decided by majority annotator opinion classified as ”Hard”, ”Easy” or ”Moderate”

- **split**: ”Train”, ”Test” or ”Val” partition, whichever the sample belongs to

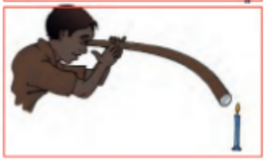
A.4 Additional Dataset Samples

We provide more examples from the VLQA dataset to visualize the diversity offered by the corpus and importance of joint reasoning to derive conclusions for real-world scenarios.

Image(s) (I)

[0]

[1]



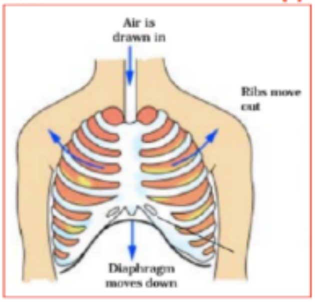
Text Passage (P) One can see the candle through [0] but not through [1].

Question (Q) Which scientific phenomena best supports the passage?

Answer Choices (A)
a. Cardboard box absorbs light rays before it reaches to candle.
b. **Light always travels in straight line**
c. Light rays can bend.
d. Cardboard box reflects light rays before it reaches to candle.

Image(s) (I)

[0]



Text Passage (P) [0] demonstrates the flow of air in Inhalation process when we breathe. Inhalation and Exhalation are complementary processes in breathing.

Question (Q) Which of the following is not correct about Exhalation?

Answer Choices (A)
a. Ribs move inside.
b. Diaphragm moves up.
c. Air is drawn out.
d. **Ribs move outside.**

Image(s) (I)

[0]

Album	Size
Album 1	100 MB
Album 2	75 MB
Album 3	80 MB
Album 4	55 MB
Album 5	60 MB
Album 6	80 MB
Album 7	75 MB
Album 8	125 MB

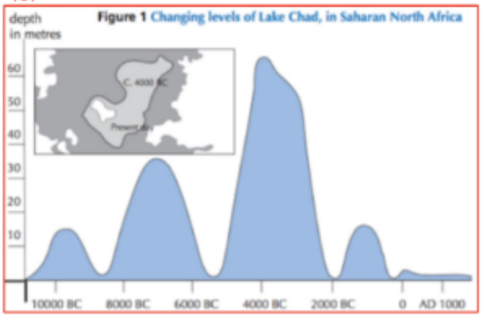
Text Passage (P) Noel's disk has 112 MB free space as of now. But he wants to store a photo album worth 350MB.

Question (Q) Can he make the space for the photo-album by deleting at most two music albums from information given in [0]?

Answer Choices (A)
a. Yes
b. No

Image(s) (I)

[0]



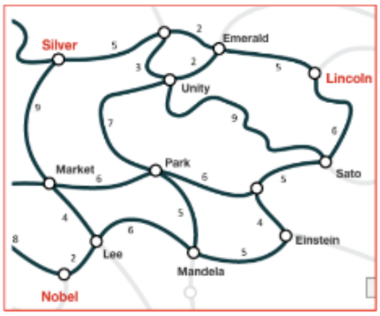
Text Passage (P) Figure [0] shows the changing levels of Lake Chad in Saharan North Africa. Today, its level is about the same as it was in AD 1000.

Question (Q) What is the depth of Lake Chad today?

Answer Choices (A)
a. **About 2m**
b. About 15m
c. About 50m
d. It disappeared completely

Image(s) (I)

[0]



Text Passage (P) Figure [0] represents the abstract map of a city, where circles represent the subway stations and the edge connecting them represents the average travel time between two stations. Julio, Maria and Don plans to meet-up at one of the subway stations today and they live in Silver, Lincoln and Nobel respectively. However, no one wants to travel for more than 15 minutes.

Question (Q) Which subway station Julio, Maria and Don could meet based on the provided information in the passage?

Answer Choices (A)
a. **Park** b. Unity
c. Market d. Emerald

Image(s) (I)

[0]

[1]

[2]

[3]

[4]

Text Passage (P) Alice and Bob are playing a game where turn-by-turn a person removes a block from the table. Starting from configuration in [0], Bob takes the first turn and removes the Purple block.

Question (Q) Choose the correct image [1] to [4] that describes configuration after the first move by Bob.

Answer Choices (A) a. [1] b. [2] c. [4] d. **[3]**

Additional Dataset Samples- Continued

Image(s) (I)

[0]

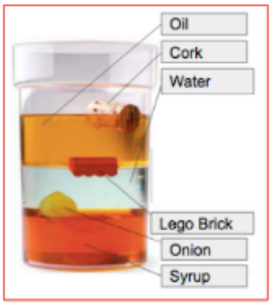
Text Passage (P)
Rescue Helicopters use the concept of Balanced forces to stabilize on water as shown in [0]. Forces in opposite directions can be Balanced out.

Question (Q)
Which of the following is a correct pair of balanced forces for a rescue helicopter?

Answer Choices (A)
a. lift and thrust
b. drag and gravity
c. **lift and gravity**
d. friction and gravity

Image(s) (I)

[0]



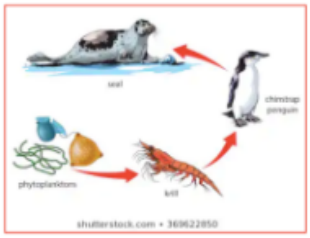
Text Passage (P)
Objects and liquids float on liquids of higher density and sink through liquids of lower density.

Question (Q)
Based on [0], which of the following is true for density of water d(W) and density of Lego brick d(B)?

Answer Choices (A)
a. **d(W) > d(B)**
b. d(W) < d(B)
c. d(W) = d(B)
d. Cannot be answered from [0]

Image(s) (I)

[0]



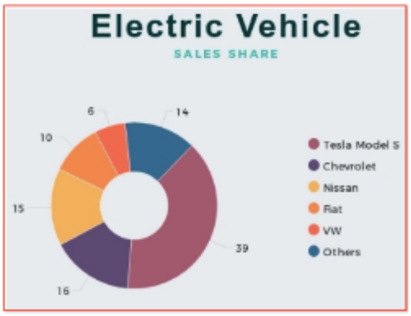
Text Passage (P)
In the recent report, World Eildlife Fund (WWF) declared Chinstrap Penguin as an endangered animal and predicted that it might extinct in near future.

Question (Q)
Which species in the aquatic food web shown in [0] most likely to be affected by outcome of this report?

Answer Choices (A)
a. Krill
b. Phytoplanktons
c. **Seal**
d. None

Image(s) (I)

[0]



Text Passage (P)
Figure [0] represents the electric vehicle sales share of different companies. A blog published on Tesla's website confirm that Tesla has officially acquired VW.

Question (Q)
How much share of the electric vehicles market will be dominated by Tesla after this acquisition?

Answer Choices (A)
a. 40% b. 55%
c. **45%** d. 50%

Image(s) (I)

[0]

City	Date opened	Kilometres of route	Passengers per year (in millions)
London	1863	394	775
Paris	1900	199	1191
Tokyo	1927	155	1928
Washington DC	1976	126	144
Kyoto	1981	11	45
Los Angeles	2001	28	50

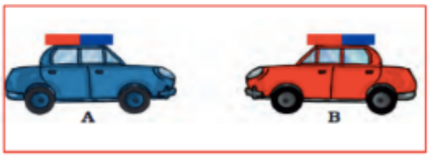
Text Passage (P)
[0] contains information about the railway network in various cities. Tokyo and Kyoto are Asian cities. London and Paris are in Europe. Los Angeles and Washington DC are popular American cities.

Question (Q)
Which continent has the smallest total railway route?

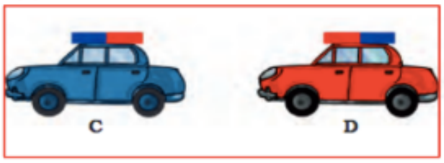
Answer Choices (A)
a. Asia b. Africa
c. **North America** d. Europe

Image(s) (I)

[0]



[1]



Text Passage (P)
Consider two cars carrying large magnets on top of them as per [0] and [1]. Red indicates north pole whereas blue indicates south pole. When you place the north pole of one magnet near the south pole of another magnet, they are attracted to one another. When you place like poles of two magnets near each other, they repel.

Question (Q)
Choose the correct statement about [0] and [1] based on the above information.

Answer Choices (A)
a. [0] will repel, [1] will attract b. **[0] will attract, [1] will repel**
c. both [0] and [1] will attract d. both [0] and [1] will repel

B Supplemental Material

Computing Infrastructure All experiments are done over Tesla V100-PCIE-16GB GPU.

B.1 Converting Visual COPA dataset into Image-Image Entailment Task

VCOPA dataset contains visual questions with three images- one labelled as premise (P) image, and other two as alternatives (H1 & H2). The task is to identify plausible alternative image related to the premise. We convert VCOPA item into Image-Image entailment task as 2-way classification as below;
Given VCOPA sample:
<P, H1, H2> | label: H1 (i.e. plausible choice)
Converted Image-Image Entailment samples:
<P, H1> | label: Entailment
<P, H2> | label: Contradiction
Then a custom 3-layer network is trained to maximize the above classification

B.2 Converting Visual COPA dataset into Text-Image Entailment Task

Similar to above, we convert VCOPA item into Text-Image entailment with additional Image Captioning module.
Given VCOPA sample:
<P, H1, H2> | label: H1 (i.e. plausible choice)
Using the Image Captioning module, get a caption for P i.e. C_P, while keeping H1 and H2 in the image format itself. Converted Text-Image Entailment samples:
<C_P, H1> | label: Entailment
<C_P, H2> | label: Contradiction
Then a custom 3-layer network is trained to maximize the above classification.

B.3 Model Parameters

Detailed summary of various components implemented for this paper - Brief description, Reference Code Link and Parameters provided in Table 6.

	Usage	Ref.	Setting
Quality Check			
Q-only: RoBERTA	RoBERTa large + RACE Ft. + ARC Ft. - Predict on VLQA <Q,A>	Link	RoBERTa Large ft. on RACE with LR=1e-5, BS=16, WD=0.1, LRD=Linear, EP=4, WR=0.06 Further ft. on ARC with BS = 8, EP = 4, LR = 1e-5
P-only: ALBERT	ALBERT-xxl + RACE Ft. - Predict on VLQA <P,Q,A>	Link	ALBERT-xxl (v2) ft. on RACE with LR=1e-5, BS=32, DR=0
I-only: LXMERT	LXMERT + VQA Ft. - Predict on VLQA <I,Q,A>	Link	LXMERT ft. on VQA with BS=32, LR=5e-5, EP=4
Pre-trained VL			
VL-BERT	VL-BERT + VQA Ft. + VLQA Ft. <I,P+Q,A>	Link	VL-BERT ft. on VQA with EP=20, BS=256, LR=1e-4, WD=1e-4 Ft. on VLQA with BS=16, LR=2e-5, EP=15
VisualBERT	VisualBERT + VQA Ft. + VLQA Ft. <I,P+Q,A>	Link	VisualBERT ft. on VQA with BS=32, LR=2e-5, EP=10 Ft. on VLQA with BS=16, LR=1e-5, EP=10
ViLBERT	ViLBERT + VQA Ft. + VLQA Ft. <I,P+Q,A>	Link	ViLBERT ft. on VQA with BS=32, LR=1e-5, EP=20, WR=0.1 Ft. on VLQA with BS=32, LR=1e-5, EP=10
LXMERT	LXMERT + VQA Ft. + VLQA Ft. <I,P+Q,A>	Link	LXMERT ft. on VQA with BS=32, LR=5e-5, EP=4 Ft. on VLQA with BS=16, LR=5e-5, EP=8
Proposed HOLE			
Text Entailment	RoBERTa large + MNLI Ft.	Link	RoBERTa Large ft. on MNLI with LR=1e-5, BS=16, WD=0.1, LRD=Linear, EP=10, WR=0.06
I-T Entailment	Bilateral Multi-Perspective Matching (BiMPM) on SNLI	Link	–
I-I Entailment	VCOPA dataset converted into Image-Image Entailment Trained custom 3-layer network for 2-class classification	Link	LR=1e-5, BS=16, OP=Adam, WD=0.1, EP=10
T-I Entailment	VCOPA dataset converted into Text-Image Entailment + Captioning Trained custom 3-layer network for 2-class classification	Link	LR=1e-5, BS=16, OP=Adam, WD=0.1, EP=10
LXMERT	LXMERT + VQA Ft. + VLQA Ft.	Link	LXMERT ft. on VQA with BS=32, LR=5e-5, EP=4
ALBERT+LXMERT	ALBERT-xxl + RACE Ft. - Predict on VLQA <P,Q> to get A' Generate Q' as "Where is A'?" and substitute A' with above string LXMERT + VQA Ft.. - Predict on VLQA <I,Q',A¿	Link Link	ALBERT-xxl (v2) ft. on RACE with LR=1e-5, BS=32, DR=0 LXMERT ft. on VQA with BS=32, LR=5e-5, EP=4

Table 6: Links to the reference code used for the paper and relevant parameters
BS - Batch Size, DR - Dropout Rate, EP - Epochs, LR - Learning Rate, LRD - Learning Rate Decay, WD - Weight Decay, WR - Warmup Ratio, Ft. - Manual Finetuning