# Understanding ME?
# Multimodal Evaluation for Fine-grained Visual Commonsense

**Zhecan Wang[1], Haoxuan You[1], Yicheng He[1], Wenhao Li[1], Kai-Wei Chang[2], Shih-Fu Chang[1]**
[1] Columbia University, New York, [2] University of California, Los Angeles
{hy2612, rs4110, zw2627, sc250}@columbia.edu,kwchang@cs.ucla.edu

## Abstract

Visual commonsense understanding requires Vision Language (VL) models to not only understand image and text but also cross-reference in-between to fully integrate and achieve comprehension of the visual scene described. Recently, various approaches have been developed and have achieved high performance on visual commonsense benchmarks. However, it is unclear whether the models really understand the visual scene and underlying commonsense knowledge due to limited evaluation data resources. To provide an in-depth analysis, we present a Multimodal Evaluation (ME) pipeline to automatically generate question-answer pairs to test models' understanding of the visual scene, text, and related knowledge. We then take a step further to show that training with the ME data boosts model's performance in standard VCR evaluation. Lastly, our in-depth analysis and comparison reveal interesting findings: (1) semantically low-level information can assist learning of high-level information but not the opposite; (2) visual information is generally under utilization compared with text.

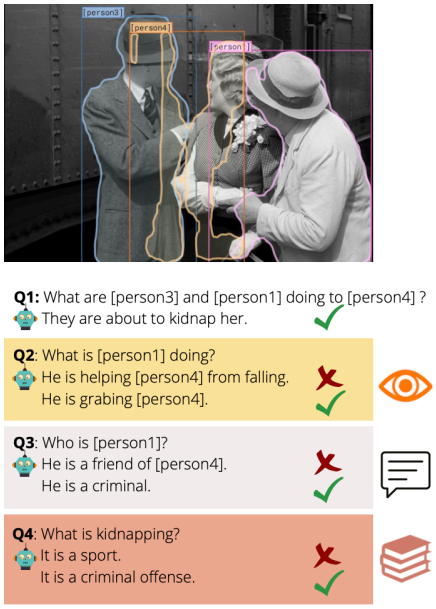


Figure 1: An example from VL benchmark, Visual Commonsense Reasoning (VCR) (Zellers et al., 2019). VL models can answer the highly semantic VCR question correctly but fail terribly in answering related visual question (Q2), textual question (Q3), and background knowledge question (Q4).

## 1 Introduction

Vision Language (VL) understanding is challenging because it requires VL models to identify and integrate information from both modalities to fully understand visual scenes. Numerous VL benchmarks have been created such as CLEVER (Johnson et al., 2017), GQA (Hudson and Manning, 2019), VQA (Li et al., 2018), VCR (Zellers et al., 2019) and SNLI-VE (Xie et al., 2018). These benchmarks typically form VL evaluation in question-answering format with images and test models' understanding of both VL modalities. Despite the high accuracy achieved by existing large pretrained VL models, recent works have pointed out that VL models tend to exploit data biases such as shallow mappings with language priors and unbalanced utilization of information between modalities (Cao et al., 2020; Jimenez et al., 2022; Selvaraju et al., 2020).

As in Fig. 1, notwithstanding the model's success in answering Q1, the same model fails to answer the related visual, textual, and background questions. This example demonstrates that the VL model does not fully understand the visual scene, which leads to prediction inconsistency (when one model makes conflicting(inconsistent) predictions in two related questions). In our analysis, prediction inconsistency is surprisingly common among models in different modalities. Former works have also pointed out that most VQA systems achieve only middling self-consistency ($60-80\%$)



*Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9212–9224
December 7-11, 2022 

(Jimenez et al., 2022). Therefore, we pose doubts to existing VL models' ability to thoroughly comprehend visual commonsense despite their high accuracy performance on the leaderboards.

In this work, we propose to evaluate models' understanding and consistency in predictions across modalities. For that intention, we propose a Multimodal Evaluation (ME) evaluation schema that can augment existing VL benchmarks like VCR. For any given sample of the VL data, *e.g.* image-question-answer pair, ME first retrieves and extracts related information of three modalities: vision, text, and background knowledge. After that, it unifies the information across modalities via a multimodal graph and further automatically generates related sub-questions corresponding to all three modalities (as examples shown in Fig. 1).

The sub-questions would be semantically relevant to the input image-question-answer pair, and therefore, after answering the original input question, we can further utilize them to evaluate existing VL models' understanding across the three modalities and pinpoint their shortcoming and biases. Under minimal human verification, with ME, we create Multimodal Sub-question Evaluation Benchmark with 630k multiple choice sub-questions for 110k images from VCR (Zellers et al., 2019): 110k of them are visual; 260k of them are about text; and the rest 260k are related to background knowledge.

After in-depth evaluation and analysis with top-performing VL models, we discover a few interesting findings: (1) semantically low-level information can assist learning of high-level information but not the opposite; (2) visual information is generally under utilization compared with text. (3) VL models may struggle to utilize related background knowledge information.

Besides, we propose a Multimodal Coaching (MC) framework to conditionally augment sub-questions in training. Depending on VL models' behavior, MC would conditionally decide if it should augment to reinforce the understanding of a particular modality. We show that by using MC, we not only improve models' consistency but also the overall performance. For example, MC boosts the performance of VL-BERT by more than $1\%$ on the original VCR Q2A metric and even more then $7\%$ in sub-question evaluation metric.

Our contributions include:

1. We identify while existing VL models perform well in commonsense benchmark, they cannot answer related sub-questions.
2. Our proposed fine-grained automatic evaluation approach allows the communities to better evaluate VL models. The code/dataset will be released upon acceptance.
3. Our in-depth evaluation and analysis with top-performing VL models discover that: (1) Training with semantically low-level information may assist learning high-level concept but not the opposite; (2) Visual information is generally under utilized compared to textual information.

## 2 Related Work

Biases occur if VL models cannot comprehensively understand the contents of both images and texts. They need to not only understand information from the two modalities respectively but also integrate these information by cross-referencing. (Cao et al., 2020; Manjunatha et al., 2019) pointed out biases like unbalanced utilization between visual and textual information. Based on these findings, previous works proposed different methods for countering them in VL benchmarks. For instance, (Agrawal et al., 2018; Zhang et al., 2016; Dancette et al., 2021) diversifies and shifts VQA's answer distribution (Goyal et al., 2017) to balance the dataset; (Gokhale et al., 2020; Liang et al., 2020a; Gupta et al., 2022; Liang et al., 2020b) augments images or creates counterfactual images to train more robust models on VQA; (Niu and Zhang, 2021; Ramakrishnan et al., 2018; Niu et al., 2021; Wang et al., 2022; Zhang et al., 2021) regularizes models' training with prior knowledge to avoid learning biases; (Ye and Kovashka, 2021) directly aligns pronouns to demonstrate biases in VCR (Zellers et al., 2019), *etc.* However, none of them helps us evaluate VL models' understanding on each modality. Without understanding how much models understand the image, the text, or the background knowledge, it is difficult to further regularize models in training.

Recently, large pretrained VL models, which are mostly trained as implicit black boxes, have been dominating VL benchmarks. It is difficult to know if they understand the image and the textual information other than simply memorizing it. Question-answering is the most general format for evaluating a wide range of models while having minimal requirements. (Ray et al., 2019; Ribeiro et al., 2019; Selvaraju et al., 2020) annotated addi-

tional questions on top of VQA questions to measure VL models' consistency on prediction. However, these works only focus on semantically low-level dataset like VQA and do not apply to highly semantic dataset like VCR. Moreover, their data fully relies on manual annotation and thus is hard to scale. Furthermore, they also fail to evaluate models' understanding across modalities. To solve these problems, we create a VL evaluation method that generates data with minimal human efforts, differentiates evaluation between modalities, and applies to highly semantic VL benchmarks. [1]

## 3 Multimodal-Eval (ME)

Given an input image-question(-answer) pair, ME first analyzes information from the pair. Then it would generate three follow-up fine-grained questions called sub-questions corresponding to three modalities: vision, text, and background knowledge. Following VCR's format, each sub-question also has four answer choices with one correct answer. The VL models are expected to first answer the VCR question and then answer the three related sub-questions. Through evaluating predictions of the sub-questions, we can test models' understanding across modalities. Overall, ME has two parts: (1) **Multimodal QA Generator** that generates related sub-questions of three modalities and (2) **Evaluation**, which test VL models' capabilities with the sub-questions. We structure the presentation as below: the method section explains the QA generation process and the experiment section discusses the evaluation process.

## 4 Multimodal QA Generator

We introduce Multimodal QA Generator through the following steps in order: (1) Retrieving related sentence statements of three modalities against the input image-text, (2) Parsing the statements into three unimodal graphs and then merging them into a multimodal graph, (3) Converting triplets in a multimodal graph into question and answer, (4) Distractor Generation, (5) Adversarial Filtering.

### 4.1 Retrieving Statements

For producing relevant sub-questions, we need to first analyze the input image-text pair and even extract information from it. Therefore, it is intuitive to have the input image and text information represented in the same level of complexity. Because the input question-answer is already in text format, we want to convert the image into text.

**Visual Statement:** Most of the existing highly-semantic VL benchmarks build on top of image/video captioning dataset, *e.g.* VQA (Goyal et al., 2017) from COCO Captions (Chen et al., 2015), SNLI-VE (Xie et al., 2018) from Flickr30K (Young et al., 2014), VCR from LSMDC (Rohrbach et al., 2017), *etc.* Those captions are visually descriptive and are not included in the image-question-answer pair. Therefore, we can directly retrieve those already annotated captions as visual statement.

**Textual Statement:** The input text prompt, *e.g.* the question-answer pairs in VCR, can be converted into statements with heuristic templates. For instance, the QA in Fig. 2 can be converted to "Person1 plays a trombone in front of everyone" + "because" + "he is performing a solo". We can also regard the converted statement as textual statement, as shown in Fig. 2 (Details in Appendix).

**Background Knowledge Statement:** In order to obtain background knowledge relevant to the visual scene, we apply keyword extractors (Campos et al., 2020) to extract keywords from visual and textual statements. Then we can regard those keywords as query concepts (as illustrated in Fig. 2, query concepts "trombone" and "solo" are extracted from the visual and textual statements). Based on query concepts, we can further browse external knowledge database, *i.e.* ConceptNet (Speer et al., 2017) to retrieve 1-hop related concepts[2] through a pool of hand-selected relationships[3]. As illustrated in background knowledge graph in Fig. 2, different triplets consisted of (Subject, Predicate, Object) are retrieved and can be conveniently converted into basic Subject-Verb-Object (SVO) sentences.

### 4.2 Generating Graph

For better integrating information, we leverage a language parser to parse the statements so that we can obtain semantic roles: Subject, Predicate and Object (S, P, O). These roles help further unify fine-grained information across three modalities. With

[1] This paper mainly focuses on applying ME on VCR, but our method can also be applied to other VL dataset consisted by image-text pairs. We also have tried it on Visual Question Answering (VQA) (Li et al., 2018) and Visual Entailment (SNLI-VE) dataset (Xie et al., 2018). Details in Appendix.

[2] https://github.com/ldtoolkit/conceptnet-lite

[3] PartOf, IsA, HasSubevent, Synonym, Antonym, MadeOf, DerivedFrom, DefinedAs, RelatedTo, UsedFor, CapableOf, AtLocation, Causes, HasProperty, Desires, CreatedBy, DistinctFrom, SymbolOf, LocatedNear, SimilarTo

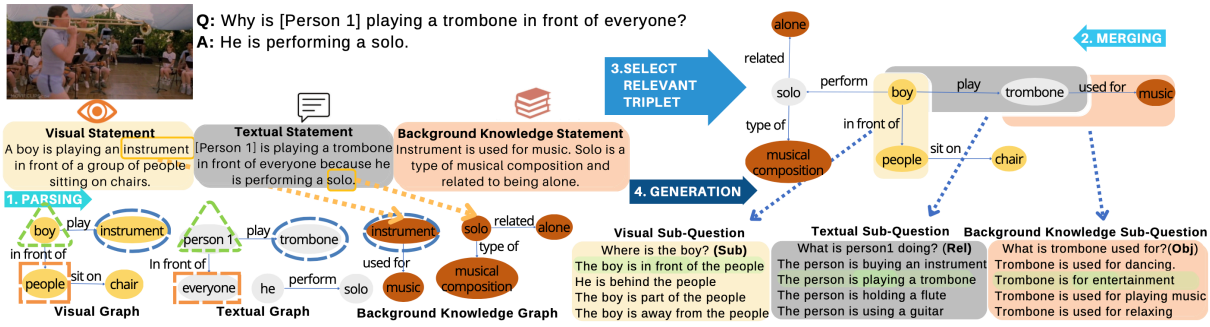


Figure 2: Pipeline of Multimodal Question-Answer Generation.

comparing them, we may find connections.

**Domain-specific Graph:** The background knowledge we retrieved from ConceptNet is already in a graph consisted of triplets (S, P, O). Therefore, we only apply the Scene Graph Parser (Schuster et al., 2015) to parse visual and textual statements into graphs. As shown in Fig. 2, we then have three domain-specific graphs corresponding to all three modalities.

**Multimodal Graphs:** To merge two graphs, we take turns to compare the similarity between each pair of nodes from them. During comparison, we not only measure their concepts' similarity but also their neighbors/context similarity. If they are similar, they would be merged into one node. For instance, given graph $G_1$ and $G_2$, we compare every node $v_i, i \in [0, \ldots, n]$ in $G_1$ with every node $v_j, j \in [0, \ldots, m]$ in $G_2$. We calculate the semantic similarity score between them, $sim_c\left(v_i, v_j\right)$ through an external tool (Zhu and Iglesias, 2017)[4].

Subsequently, we also compare the neighbors of $v_i$ against the neighbors of $v_j$. Let's assume $v_i$ has $p$ 1-hop connections in $G_1$ and $v_j$ has $q$ 1-hop connections in $G_2$. Every connection of $v_i$ links two concepts thus forming a triplet. It can be converted into a SVO sentence containing $v_i$ as either the Subject or Object. With that, we result in $p$ sentences related to $v_i$. Similarly, we can also obtain $q$ sentences related to $v_j$. Following (Ni et al., 2022), we inference a pretrained Sentence-T5 to extract $p$ and $q$ sentence embeddings. Then, for every pair between $S_l, l \in [0, \ldots, p]$ and $S_o, o \in [0, \ldots, q]$, we calculate the cosine distance $sim_s\left(S_l, S_o\right)$. Lastly, for every pair, $(v_i, v_j)$, we sum both node concept similarity and context similarity together by:

$$\text{Score}_{node}\left(v_i, v_j\right) = \text{sim}_c\left(v_i, v_j\right) + \frac{\sum_o \sum_l \text{sim}_s\left(S_l, S_o\right)}{p \cdot q}. \quad (1)$$

If $\text{Score}_{node}\left(v_i, v_j\right)$ is larger than a threshold $T$ (Details in Appendix), we would consider $v_i, v_j$ as duplicates and only keep one in the graph.

**Selecting Relevant Sub-graphs:** After obtaining the graph representation, we want to generate sub-questions relevant to the input image and VCR question of each sample $u$. Therefore, we filter each triplet in the multimodal graph by its relevance against the input image-question-answer pair.

Similar to above, we convert all $r$ triplets in multimodal graph into sentences $S_k^u, k \in [0, \ldots, r]$ and measure their similarity to the textual statement (a conversion of the input QA) via (Ni et al., 2022), $sim_s\left(S_k^u, S_{QA}^u\right)$.

Afterwards, we further utilize a pretrained CLIP (Radford et al., 2021) to encode and then calculate the cosine distance between every sentence against the image $I^u$, $rel_s\left(S_k^u, I^u\right)$. In conclusion, the final score for every triplet would be:

$$\left\|\text{sim}_s\left(S_k^u, S_{QA}^u\right)\right\|_1 + \left\|rel_s\left(S_k^u, I_u\right)\right\|_1. \quad (2)$$

After ranking, we select the top-1 ranked triplet in every modality. If a triplet is selected in more than one modality, we replace the duplicate with the next following triplet in the same modality.

### 4.3 QA Templates

Given a triplet, we can ask questions about the subject, the object, or the predicate. For instance, in (boy, in front of, people), if asking about the object, we could use templates like "What is the [Subject] [Predicate]" (What is the boy in front of?). In this case, the basic answer would be [Object](People) or the converted full SVO sentence of

[4]It measures the distance of the two nodes' concepts in WordNet (Ingo Feinerer, 2020) and YAGO (Pellissier Tanon et al., 2020) and then averages the reverse of the two distances as the similarity score

the triplet, [Subject][Predicate][Object](The boy is in front of people). When asking about the subject or the predicate, similar procedure applies (Details in Appendix)

### 4.4 Distractor Generation

The new evaluation task should have the same format as the original one (multiple-choice-format (MCQ) in VCR) so that we can directly evaluate existing models. For that purpose, it is necessary to generate incorrect answer choices, distractors.

Simply rephrasing the correct answer may produce false negative that confuse the models. In a sense, more non-trivial and meaningful disturbance should be added to the answer distribution.

In practice, we choose to represent the answer in SVO sentence format, *e.g.* "The boy is in front of people". We first parse the answer into (Subject, Predicate, Object) *e.g.* (boy, in front of, people) and regard this as the starting templates for creating distractors. If the question is asking about the relationship, then we could regard relationship as the "changeable part" in the template. We could replace this "changeable part" with other words to create new combinations for distractors *e.g.* (boy, behind, people). In order to make meaningful replacement, we use the original relationship concept, "in front of", as the query concepts to retrieve related concepts from external resources like "behind", "direction", "location". We apply the same procedure to the subject and object.

**Explicit Retrieval from External Knowledge:** We follow a similar procedure in retrieving background knowledge concepts from ConceptNet (Speer et al., 2017), while only differing in our selection of a different set of relationships (Details in the Supple).

**Implicit Retrieval from Language Models:** We also utilize pretrained language models to help retrieve related concepts in two perspectives. First, in cases when the question is asking about the object and the program fails to retrieve related concepts from explicit resources, we leverage prompt engineering to implicitly retrieve related concepts from a pretrained language model, GPT2 (Radford et al., 2019) alternatively. Using the same triplet as an example, if the question asks about the object, then we would design the prompt as "boy is in front of [mask]". After GPT2 fills in the [MASK], we should be able to retrieve external concepts within GPT2's top predictions. We can further use as them as options for objects in distractors.

After successfully replacing concepts in the template, we directly apply heuristic rules and convert it into SVO sentences with heuristic rules to create a distractor. Aiming for variety beyond rule-based sentence construction, we also alternatively use another language model to process the conversion. We exploit a sentence generation model, T5 fine-tuned on CommonGen (Lin et al., 2020) which is built on top of ConceptNet. The training task in CommonGen is to convert set of concpet words into everyday sentences. For example, after replacing concepts in the template, from *(boy, in front of, people)* to *(boy, back, people)* and *(boy, direction, people)*, we input them directly into T5, which outputs a list of possible sentences *e.g.* "A boy is running back to the people", "A boy is facing the same direction as the other people". Different from hard-coded templates used to generate SVO sentences, T5 fills in context words around the input concepts, thus also helps retrieving implicit external concepts like ("running", "facing", "same", "other"). These additional concepts may not be relevant to the visual scene which aligns with the purpose of generating distractors.

### 4.5 Adversarial Filtering

High-quality distractors should be semantically related to the answer but also different enough for humans to tell. Therefore, we design our own adversarial filtering (Zellers et al., 2018, 2019) mechanism by using pretrained VL and language models to filter data. We first correct all generated distractors by an off-shelf grammar checker[5]. Then we further filter them by a pretrained language model to remove distractors that are too semantically close to the correct answer to reduce potential false negatives. Lastly, we apply a pretrained VL model to measure their relevance against the image and select the top three as final distractors (Details in Appendix).

## 5 Dataset

**Dataset Statistics** Built on top of (Zellers et al., 2019), Multimodal Sub-question Evaluation Benchmark has around 110k visual sub-questions corresponding to the 110k images from (Zellers et al., 2019), 260k text(prompt) sub-questions, and 260k background knowledge sub-questions corresponding to the 290k original questions from (Zellers

[5]https://pypi.org/project/language-tool-python

| Metric | Generated | Verified |
|---|---|---|
| Individual Acc. | 0.83 | 0.89 |
| Group Acc. | 0.95 | 0.99 |
| Group Top2 Recall | 0.94 | 0.98 |
| IAA | 0.82 | 0.88 |

Table 1: Comparison between generated and verified data. Every sample has five annotations/selections. Individual Acc. represents the accuracy when each annotator's selection is counted as one prediction. Group Acc. represents the accuracy when only the highest frequent selection is counted as the prediction for the group. Group Top2 Recall represents the accuracy if the groundtruth is within the top 2 most frequent selections of the group. IAA is the Inter-Annotator Agreement

et al., 2019)[6]. Every question has four answer choices and the answers have an average length of 5.5 words. The ratio between training set and validation set is $10:1$.

**Quality Control** To deliver a convincing evaluation method to existing VL models, we have humans verify the full validation set. We designed and deployed a user interface on Amazon Mechanical Turk platform and hired experienced turkers (with $\$12.6/hr$)to help verify the correctness of our questions and answers. Every image-question pair was cross-verified and corrected by five turkers (Details in Appendix).

**Evaluation** We randomly select 2 disjoint sets each containing 100 image-question pairs from ME. The first set consists generated QA data. The second one consists QA data verified and corrected by the turkers. We then hire an additional group of five turkers to answer those 200 image-question pairs without knowing the answer label. Next, we calculate the predictions' accuracy. As in Tab. 1, the difference between generated and verified data is very minimal which demonstrates the high-quality of our generated data (More in Appendix).

## 6 Evaluation

In the following, we conduct experiments based on the proposed dataset to demonstrate (1) The existing models that perform well on VL dataset often cannot answer detailed vision, text, knowledge sub-question correctly; (2) It is easier for VL models to answer sub-questions originated from semantically low-level VCR questions than high-level ones.

[6]One image has only one visual sub-question but may correspond to multiple text or background knowledge sub-questions. Some of the original VCR questions are too short and do not contain meaningful sub-questions

**Base Methods** During our experiments, we use three top-performing models, VL-BERT (Su et al., 2019), UNITER (Chen et al., 2020) and VILLA (Gan et al., 2020) on VCR leaderboard as our base models.

**Evaluation Metrics** When calculating the original Q2A accuracy of VCR (Zellers et al., 2019) on $n$ total samples, let $C_j^{Q2A}$ be an indicator variable for sample $j, j \in [0, \ldots, n]$. If the prediction, $P_j^{Q2A}$, is the same as the label, $L_j^{Q2A}$, the prediction is correct and $C_j^{Q2A} = 1$; otherwise $C_j^{Q2A} = 0$.

$$\text{Accu.}_{\text{Q2A}} = \frac{\sum_{j=0}^{j=n} \mathrm{C}_j^{Q2A}}{n}.$$

Similarly, in our new metrics, we have indicator variables $\mathrm{C}_j^{Q2S-x}$ for the correctness of prediction on sub-questions related to modality $x$, which can be vision, text, or background knowledge; similarly, we use $\mathrm{C}_j^{Q2AS-x}$ to indicate the event that both the VCR question and the sub-question corresponding to modality $x$ are correct. Lastly, $\mathrm{C}_j^{Q2S}$ indicates the event that all the sub-questions related to sample $j$ are predicted correctly.

$$\begin{cases} \mathrm{C}_j^{Q2S-x} = 1, \text{ if } P_j^{Q2S-x} = L_j^{Q2S-x}, \text{ else } 0 \\ \mathrm{C}_j^{Q2AS-x} = 1, \text{ if } \mathrm{C}_j^{Q2A} = 1 \text{ and } \mathrm{C}_j^{Q2S-x} = 1 \\ \mathrm{C}_j^{Q2S} = 1, \text{ if } \sum \mathrm{C}_j^{Q2S-x} = 3, x \in \{V, T, BK\} \end{cases}$$

### 6.1 Comparison across Modalities

We want to evaluate VL models' capability in understanding fine-grained information from different modalities. Tab. 2 shows the evaluation results of the models. Looking at rows marked with "N" under the column name "ME in Training", we discover that existing VL models all suffer around a $20\%$ drop in accuracy on our sub-questions' metrics. Among modalities, VL models generally perform the best in textual sub-questions. This is expected since the semantic contents of the textual sub-questions are the closest one to the original VCR questions. In contrast, these models often perform slightly worse in visual sub-questions. This re-verifies previous works' concerns that existing VL models generally under-utilize visual information. Lastly, they all suffer the most in answering background knowledge sub-questions. We believe that despite background knowledge may be useful in humans' perspective, VL models are still lack of sufficient abilities to utilize them. In fact, they seem to have the largest domain gap against the

original VCR questions to VL models. These evaluation results help verify our previous hypothesis. **Consistency:** When considering consistency, even larger drops about $40\%$ occur across models' performances. The general trend of performances on Q2AS-x between modalities is similar to Q2S-x as discussed before, but with lower overall values.

### 6.2 Comparison across Question Types

The first row in Fig. 3 (A) visualizes the number of questions of each type in VCR validation set. Observing it, we can see clear imbalanced distribution exists among question types in VCR validation set. Hence, we on purpose, sample 2k questions of each type from validation set to create a balanced mini-validation set of 14k VCR image-question pairs. We further evaluate the finetuned VL-BERT on this mini-validation set. On the second row, for each question type, we visualize the number of Q2A questions that VL-BERT predicts correctly. From the third to the fifth row, we visualize sub-questions (originated from different types of Q2A questions) that VL-BERT predicts correctly. As we can see, for Q2S-V, explanation, activity and scene questions have the most percentage. Apart from the explanation and activity questions also being the most dominant question types in the training set, the semantic relatedness with the activity questions and the explanation questions also helps models answer visual sub-questions. Additionally, we realize that it is easier for the model to answer sub-questions (from all three modalities) of semantically low-level VCR questions like explanation, activity types. However it is difficult for abstract ones like mental questions.

## 7 Multimodal Coaching for Model Improvement

Besides utilizing ME data for evaluating existing VL models' performance of fine-grained understanding across modalities and prediction consistency, we also find that ME can further assist existing VL models' training.

### 7.1 Multimodal Coaching

In order to better utilize ME data and allow VL models to have a balanced learning over information from different modalities, following (Ray et al., 2019), we design the Multimodal Coaching (MC) system. According to (B) in Fig. 3, when iterating over every VCR sample in training, Multimodal QA generator produces relevant sub-questions. Then MC would take turns to test the QA model with those sub-questions across three modalities. If the model fails on any of them, the corresponding sub-question would be added to the training pool otherwise it would be passed. Therefore, we selectively augment ME data with VCR data during the training.

### 7.2 Data Augmentation

We demonstrate the effectiveness of training with ME data in Tab. 3. We keep VL-BERT as our base model and cumulatively add sub-questions across three modalities into the training set. VCR has 7 types of questions and some of them are highly semantic not much related to visual compositions like mental, hypothetical questions. With this prior knowledge, when adding visual sub-questions, we on purpose do not augment them to VCR questions of these two types.

We observe in Tab. 3 that adding visual and textual sub-questions both bring improvements on Q2A, QA2R and sub-question metrics including Q2S, Q2S-V, Q2S-T and Q2S-BK. However, adding background knowledge sub-questions hurts the performance. As mentioned before in **Evaluation** section, the additional content from external database like ConceptNet has a large domain gap against VCR questions and thus may be too difficult for VL models to utilize. However, this result further debunks existing VL models' vulnerability and confirms that it is important to include background knowledge sub-questions in VL evaluation analysis.

Lastly, looking at the last row of Tab. 3, we observe that MC could further boost VL-BERT's performance gain. In experiments, we also realize that adding MC would allow training loss to be more stable and converge faster.

### 7.3 Composite vs. Component Information

Comparing the first row against the third row in Tab. 4, we notice that VL-BERT performs better on Q2A when having both VCR questions and visual sub-questions in training set. Comparing the second row against the third row, we also discover that VL-BERT performs better on Q2S when the training set only contains visual sub-questions. Adding VCR questions would actually hurt its performance on Q2S.

We observe similar results when comparing other sets of rows like the (first, fourth, fifth) rows for text sub-questions, and the (first, sixth, seventh) rows for background knowledge sub-questions.

| Model | ME in Training | Evaluation | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | VCR | Subsequent Questions | | | | Consistency | | |
| | | Q2A | Q2S | Q2S-V | Q2S-T | Q2S-BK | Q2AS-V | Q2AS-T | Q2AS-BK |
| VL-BERT | N | 75.53 | 55.31 | 54.96 | 56.18 | 55.75 | 41.59 | 42.51 | 42.19 |
| | Y | 76.59 | 61.16 | 60.12 | 62.81 | 58.75 | 46.05 (+4.46) | 48.11 (+5.6) | 44.99 (+2.8) |
| UNITER | N | 76.64 | 57.49 | 57.83 | 57.54 | 56.34 | 44.32 | 44.1 | 43.17 |
| | Y | 77.12 | 63.51 | 62.84 | 66.04 | 60.87 | 48.46 (+4.14) | 50.93 (+6.83) | 46.94 (+3.77) |
| VILLA | N | 78.27 | 59.85 | 58.41 | 61.05 | 56.55 | 45.71 | 47.78 | 44.26 |
| | Y | 78.79 | 63.99 | 63.2 | 66.43 | 60.83 | 48.74 (+3.03) | 51.23 (+3.45) | 46.91 (+2.65) |

Table 2: Evaluation of benchmark VL models' consistency across modalities.

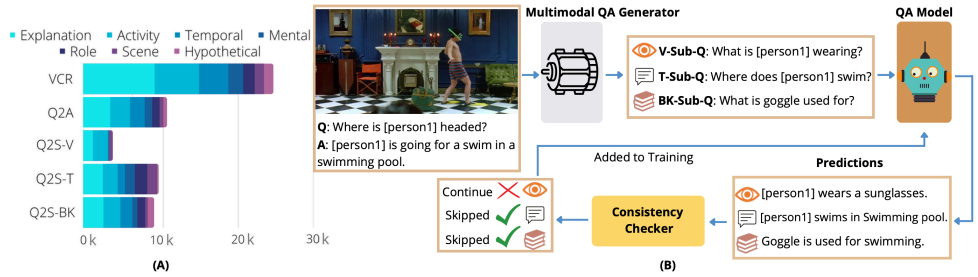


Figure 3: (A) Evaluation across Question Types. (B) Pipeline of Multimodal Coaching.

| Training | | | | | Evaluation | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| VCR | Sub-V | Sub-T | Sub-BK | MC | Q2A | QA2R | Q2S | Q2S-V | Q2S-T | Q2S-BK |
| Y | | | | | 75.67 | 77.84 | 55.31 | 54.96 | 56.18 | 55.75 |
| Y | Y | | | | 76.08 (+0.41) | 78.33 (+0.49) | 59.07 (+3.76) | 59.84 (+4.88) | 59.51 (+4.33) | 58.01 (+2.26) |
| Y | Y | Y | | | 76.59 (+0.92) | 78.76 (+0.92) | 61.16 (+5.85) | 60.12 (+5.16) | 62.81 (+6.63) | 58.75 (+3.00) |
| Y | Y | Y | Y | | 76.48 (+0.81) | 78.35 (+0.51) | 59.14 (+3.83) | 58.63 (+2.67) | 60.66 (+4.48) | 59.47 (+3.72) |
| Y | Y | Y | | Y | **76.88 (+1.21)** | **79.05 (+1.21)** | **61.89 (+6.58)** | **60.44 (+5.48)** | **63.62 (+7.44)** | **59.41 (+3.66)** |

Table 3: Data augmentation. Numbers in brackets are the difference between data in that row against the first row.

| Training | | | | Evaluation | |
|---|---|---|---|---|---|
| VCR | Sub-V | Sub-T | Sub-BK | VCR (Q2A) | Q2S |
| Y | | | | 75.67 | 55.31 |
| - | Y | | | 59.31 | 60.11 (+4.8) |
| Y | Y | | | 76.08 (+0.41) | 59.07 (+3.76) |
| - | - | Y | | 59.72 | 61.01 (+5.70) |
| Y | - | Y | | 76.20 (+0.53) | 60.33 (+5.02) |
| - | - | - | Y | 55.72 | 58.99 (+3.68) |
| Y | - | - | Y | 75.48 (-0.19) | 58.12 (+2.81) |

Table 4: Comparison between training with composite and component data. Numbers in brackets are the difference between data in that row against the first row.

Even though, when having both background knowledge sub-questions and VCR questions in training, the Q2A performance drops slightly (due to potential reasons explained above), the Q2S performance drops even much more due to adding VCR questions. Also, Q2A performance via training on background knowledge sub-questions only is even higher than the Q2S performance via training on VCR questions only(Both questions share the same MCQ format with four answer choices and random guess is 25%).

If we regard VCR questions as composite information since information from different modalities are combined together in the questions, we can then refer sub-questions as component information "parsed from" the composite information. Based on the comparison, we conclude that low-level component information could potentially help models' understanding of high-level composite information. However, after learning with high-level composite information, existing VL models may struggle to utilize the high-level to help understand low-level component information.

### 7.4 Comparison across Modalities

As in Tab. 2, after adding ME sub-question data in training, VL models generally improve in accuracy across Q2A, sub-question metrics and consistency metrics. Complementary to the findings in the **Evaluation** section, we discover that **(1)** VL models tend to have more consistent predictions in answering textual sub-questions; **(2)** Adding textual sub-questions in training also brings more improvements on sub-questions metrics corresponding to the other two modalities.

## 8 Conclusion

In this work, we propose ME to thoroughly probe VL models' understanding across and between modalities. Our analysis brings new insights and our experiments show that ME boosts models' performance when used in training.

## 9 Limitation

ME requires the given image to have paired captions so they can be easily converted into visual statements. When absent, we can inference from a pretrained caption generator at the expense of accuracy. However, sometimes the visual caption generator may not fully captures the most salient activities in the image and thus produces trivial captions with limited contents. Therefore, it would be difficult for ME to extract related information from the caption to further create the visual sub-question.

Also, technically, for any VL dataset with image-question-answer pairs, ME should be able to generate sub-questions from three modalities. However, if the input question is very simple and focuses on semantically low-level information. It woud be challenge for ME to further extract and create sub-questions from three modalities.

This study is solely based on English data and leverages linguistic structures in English so it cannot generalize to other languages.

## 10 Appendix

### 10.1 Generated vs. Verified

In Tab. 5, we evaluate VL-BERT with both generated ME data and data verified by human annotators.

In Tab. 6, we finetune a VL-BERT with both generated and verified data by humans.

Results from both tables demonstrate the high-quality of our generated data.

### 10.2 Hyper-perameter

1. In practice, the semantic similarity between concepts of two nodes would be first standardized via z-score and then compared against a hyper-parameter $T$ of 0.8.

### 10.3 Examples in other VL Benchmarks

Referring to Fig. 4, 5.

### 10.4 User Interface

Referring to Fig. 6

### 10.5 Adversarial Filtering

High-quality distractors should be semantically related to the answer but also different enough for humans to tell. Therefore, we design our own adversarial filtering (Zellers et al., 2018, 2019) mechanism by using pretrained VL and language models to filter data. We first correct all generated distractors by an off-shelf grammar checker[7]. Then we further filter them by a pretrained language model to remove distractors that are too semantically close to the correct answer to reduce potential false negatives. Lastly, we apply a pretrained VL model to measure their relevance against the image and select the top three as final distractors.

**Sentence-Similarity Modeling:** Similar to previous procedures, across all $z$ number of distractors, we compare each of them $S_w, w \in [0, \ldots, z]$ against the textual (QA) statement $S^u_{QA}$, $Score^w_{sent} = \text{sim}_s \left(S^u_w, S^u_{QA}\right)$. By removing distractors whose $Score^w_{sent}$ is above a threshold, $D$ (0.7), we reduce potential false negatives that are semantically close to the correct answer.

**Image-Text Matching:** After that, we also need to ensure that the distractors are visually relevant to the image. We load a pretrained CLIP model (Radford et al., 2021) to measure the relevance between each distractor against the image, $Rel^w_{sent} = \text{rel}_s\left(S^u_w, I^u\right)$. We rank all the distractors by $Rel^w_{sent}$ and select the top 3 distractors as the final distractors.

### 10.6 Quality Control

To deliver a convincing evaluation method to existing VL models, we have humans verify the full validation set. We designed and deployed a user interface on Amazon Mechanical Turk platform and hired experienced turkers (with $\$12.6/hr$)to help verify the correctness of our questions and answers. Every image-question pair was cross-verified and corrected by five turkers

Having the image on the side, every turker would be first asked to verify the correctness of the question in terms of grammar or understanding. If the question is marked as incorrect or not understood, we would ask the turkers to help re-correct the question or skip it[8]. Then the turkers would be provided with 7 answer choices (1 correct answer choice and 6 incorrect answer choices) and 2 additional choices of "None of the above" and "I do not know how to answer".

Avoiding causing any prior biases in the turkers and resulting in false positives and false negatives, we do not inform turkers the number of correct answer choices and ask them to select all the ones

[7] https://pypi.org/project/language-tool-python

[8] For skipped data, the author would take a look at them to verify.

| Verified | Evaluation | | | | |
|---|---|---|---|---|---|
| | VCR (Q2A) | Q2S | Q2S-V | Q2S-T | Q2S-BK |
| N | 75.67 | 54.23 | 54.81 | 54.95 | 54.87 |
| Y | 75.67 | 55.31 | 54.96 | 56.18 | 55.75 |

Table 5: Evaluation with generated and verified data.

| Verified in Training | Evaluation | | | | |
|---|---|---|---|---|---|
| | VCR (Q2A) | Q2S | Q2S-V | Q2S-T | Q2S-BK |
| N | 76.13 | 59.34 | 60.77 | 61.74 | 58.8 |
| Y | 76.59 | 61.16 | 60.12 | 62.81 | 58.75 |

Table 6: Evaluation of VL-BERT trained with generated and verified ME data augmentation.

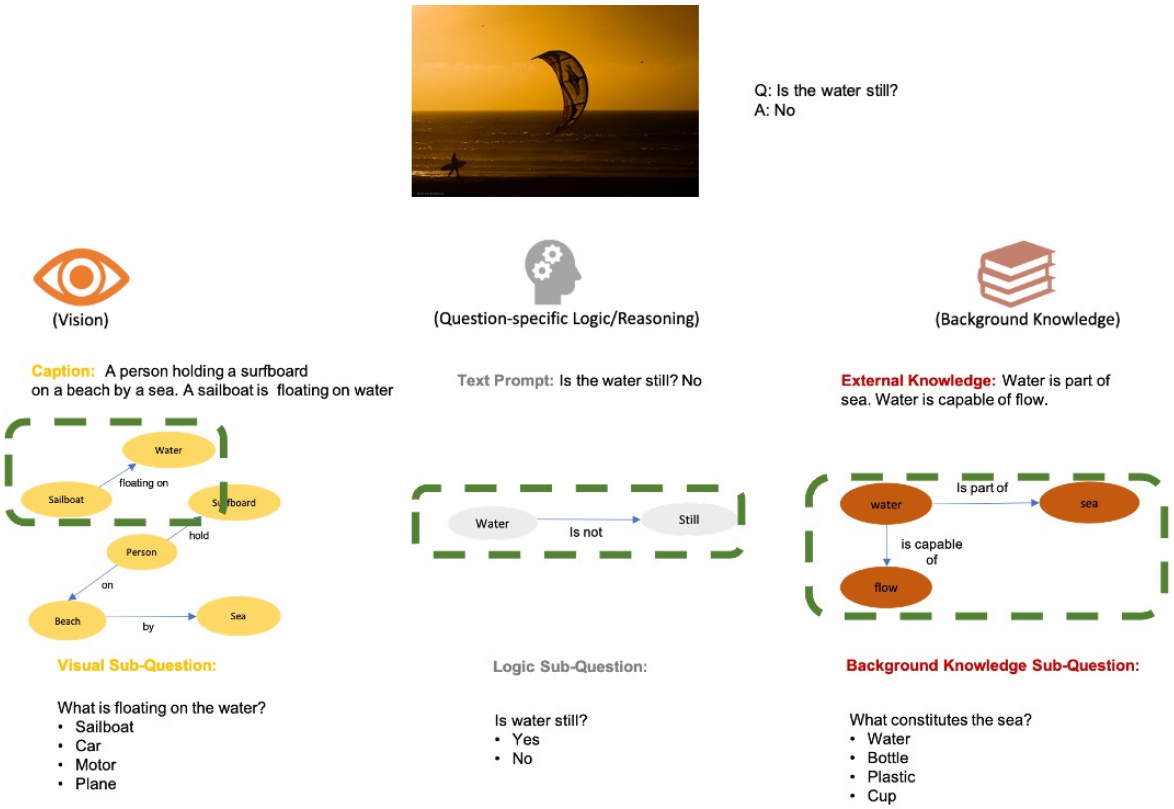


Figure 4

they think are correct. If they cannot understand the visual scene or find a correct answer at all, they can even select "I do not know how to answer" or "None of the above". After selecting the answer choices, we also give the turkers options to go over every answer choice to re-correct it if there is any grammatical issue. In the end, if the turkers have selected "None of the above" before, they would be asked to created their own correct answer choices.

To ensure the correctness of the annotation interface, we first conduct many in-house experiments. After that, we also randomly select several turkers' annotations as pseudo groundtruths. We further evaluate other turkers' annotation against the pseudo groundtruths to ensure the agreement rate on selections.

For an image-question pair, if turkers have different selections on the correct answer choices, we would avoid avoid using any answer choices selected as correct by any of the turker as a distractor.

When filtering the annotations, we ensure that every selected final distractor in ME cannot be selected by any of the turkers as correct before to avoid false negative. Further, when filtering every

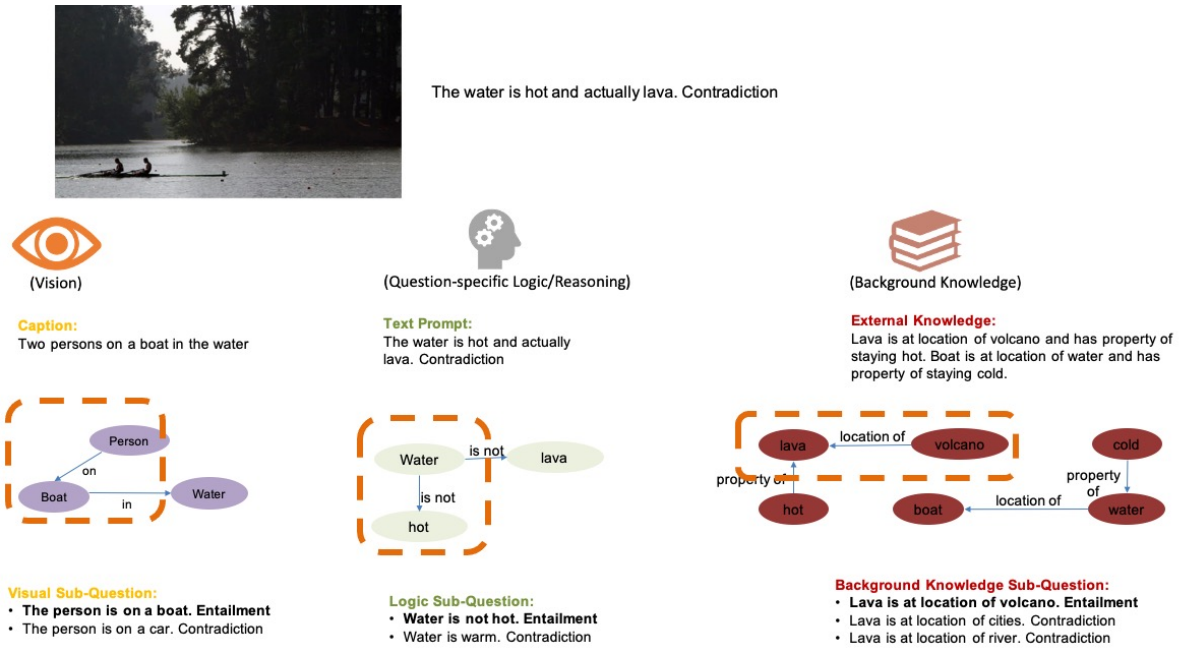


Figure 5

Figure 6: A user interface for collecting data.

sample's annotations, among the five turkers, we ensure that the selected final correct answer choice should be selected by at least three of them to avoid false positive. If more than one answer choice is selected three times, we would compare and select the one that has the most selections.

## References

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

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289.

Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. 2020. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In *European Conference on Computer Vision*, pages 565–580. Springer.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

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

Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. *Advances in Neural Information Processing Systems*, 33:6616–6628.

Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020. Mutant: A training paradigm for out-of-distribution generalization in visual question answering. *arXiv preprint arXiv:2009.08566*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Vipul Gupta, Zhuowan Li, Adam Kortylewski, Chenyu Zhang, Yingwei Li, and Alan Yuille. 2022. Swapmix: Diagnosing and regularizing the over-reliance on visual context in visual question answering. *arXiv preprint arXiv:2204.02285*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Kurt Hornik Ingo Feinerer. 2020. *wordnet: WordNet Interface*. R package version 0.1-15.

Carlos E Jimenez, Olga Russakovsky, and Karthik Narasimhan. 2022. Carets: A consistency and robustness evaluative test suite for vqa. *arXiv preprint arXiv:2203.07613*.

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

Qing Li, Qingyi Tao, Shafiq Joty, Jianfei Cai, and Jiebo Luo. 2018. Vqa-e: Explaining, elaborating, and enhancing your answers for visual questions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 552–567.

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

Zujie Liang, Weitao Jiang, Haifeng Hu, and Jiaying Zhu. 2020b. Learning to contrast the counterfactual samples for robust visual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3285–3292, Online. Association for Computational Linguistics.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, Online. Association for Computational Linguistics.

Varun Manjunatha, Nirat Saini, and Larry S Davis. 2019. Explicit bias discovery in visual question answering models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9562–9571.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. 2022. Sentence-t5: Scalable sentence encoders from pretrained text-to-text models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1864–1874, Dublin, Ireland. Association for Computational Linguistics.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12700–12710.

Yulei Niu and Hanwang Zhang. 2021. Introspective distillation for robust question answering. *Advances in Neural Information Processing Systems*, 34:16292–16304.

Thomas Pellissier Tanon, Gerhard Weikum, and Fabian Suchanek. 2020. Yago 4: A reason-able knowledge base. In *The Semantic Web*, pages 583–596, Cham. Springer International Publishing.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International*

*Conference on Machine Learning*, pages 8748–8763. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. 2018. Overcoming language priors in visual question answering with adversarial regularization. *Advances in Neural Information Processing Systems*, 31.

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

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

Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. 2017. Movie description. *International Journal of Computer Vision*, 123(1):94–120.

Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D Manning. 2015. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the fourth workshop on vision and language*, pages 70–80.

Ramprasaath R Selvaraju, Purva Tendulkar, Devi Parikh, Eric Horvitz, Marco Tulio Ribeiro, Besmira Nushi, and Ece Kamar. 2020. Squinting at vqa models: Introspecting vqa models with sub-questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10003–10011.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Zhecan Wang, Noel Codella, Yen-Chun Chen, Luowei Zhou, Xiyang Dai, Bin Xiao, Jianwei Yang, Haoxuan You, Kai-Wei Chang, Shih-fu Chang, et al. 2022. Multimodal adaptive distillation for leveraging unimodal encoders for vision-language tasks. *arXiv preprint arXiv:2204.10496*.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2018. Visual entailment task for visually-grounded language learning. *arXiv preprint arXiv:1811.10582*.

Keren Ye and Adriana Kovashka. 2021. A case study of the shortcut effects in visual commonsense reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 3181–3189.

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

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. *arXiv preprint arXiv:1808.05326*.

Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Yin and yang: Balancing and answering binary visual questions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5014–5022.

Xi Zhang, Feifei Zhang, and Changsheng Xu. 2021. Multi-level counterfactual contrast for visual commonsense reasoning. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1793–1802.

Ganggao Zhu and Carlos A Iglesias. 2017. Sematch: Semantic similarity framework for knowledge graphs. *Knowledge-Based Systems*, 130:30–32.