Alleviating the Impact of Language Bias for Evaluating Visual Commonsense Reasoning

Anonymous submission

Abstract

The VCR dataset is a large-scale visual question answering dataset, which is designed to require a high-level of image understanding and commonsense reasoning ability. However, the dataset can overestimate the visual reasoning capabilities of models as a language model can solve 70% of the samples in the dataset without the image. To handle this issue, we propose a new evaluation dataset, VCR-Visor (**VCR** benchmark dataset for **Vis**ually gr**O**unded **R**easoning), a re-configured version of the VCR dataset. VCR-Visor aims to prevent models from achieving high scores by relying on textual modality. To this end, an example of our dataset consists of two different images and corresponding answers. We have observed that the models that show high performance on VCR show much lower performance on VCR-Visor and a language model showed performance much below the random baseline. We also have discovered that increasing the model size can even result in a decline in performance, implying that larger models can be trained to rely on linguistic cues more. We conclude that there is much room for improvement of models for visual commonsense reasoning and the VCR-Visor dataset can motivate and quantitatively measure such research.

1 Introduction

Vision-Language (VL) tasks are challenging as they require models to seamlessly integrate and understand both modalities. Recently, VL tasks have gained significant attention and various benchmark datasets have been proposed (Antol et al., 2015; Johnson et al., 2017; Suhr et al., 2017; Hudson and Manning, 2019; Zellers et al., 2019; Cao et al., 2020; Zhou et al., 2022). The current most successful approaches for them are Vision-Language Pretrained (VLP) models, which have achieved outstanding performance (Du et al., 2022). Unfortunately, in cases where bias exists within the dataset, there is a tendency for models to rely on the bias in order to achieve high performance. (Agrawal et al., 2018; Cao et al., 2020; Salin et al., 2022)

Biases present in datasets can lead to an overestimation of the true performance of models, and this issue has been reported in a variety of VL tasks. Visual question answering (VQA) models depend on dataset biases such as language priors (e.g., The answer for the questions beginning with "Do you see a..." are mostly "Yes") that exist in the VQA dataset (Goyal et al., 2017; Agrawal et al., 2018). In image captioning, models tend to make biased predictions for a person's gender by using information such as location (e.g., Kitchen) (Hendricks et al., 2018). Models trained on biased datasets can make predictions that are not based on meaningful visual evidence.

We address this issue in visual commonsense reasoning (VCR). The VCR dataset (Zellers et al., 2019) is a question answering dataset for VCR that requires high-level reasoning such as inferring social dynamics, the intents of people, and the cause of the event. Similar to other datasets, there are biases in the VCR dataset. Models can achieve higher performance by selecting the answer choice which has the most overlapping referring tags (e.g., '[person1]', '[dog1]') with the question.

Including the above bias, we focus on a wide range of correlations between questions and answers in the VCR dataset. For example, consider an example of question answering with two answer candidates:

> **Q.** Why is [person2] happy?
> **A1.** She is receiving a gift.
> **A2.** [person2]'s pet seal is having a birthday party.

The language model trained on the VCR dataset somehow predicts that **A1** is a more plausible answer than **A2**. This model can find the correct answer in more than 70% of the samples in the

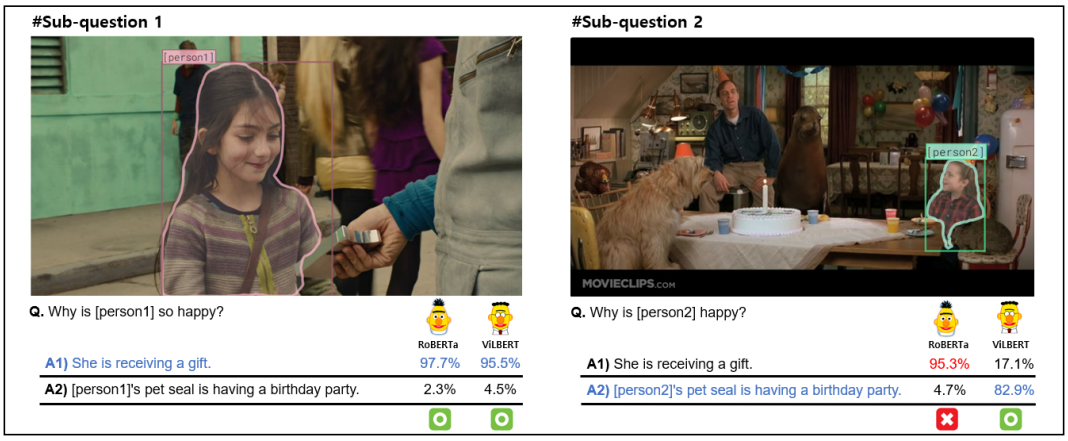


Figure 1: An example of the VCR-Visor dataset and predictions of language model RoBERTa (Liu et al., 2019) and VL model ViLBERT (Lu et al., 2019). The example consists of two sub-questions and each answer for the sub-question is highlighted in blue. Models need to leverage visual information to solve both sub-questions.

VCR dataset without the image. However, since the model lacks visual understanding, it will still choose **A1** even if an appropriate image for option **A2** is given. We investigate whether VL models also rely on the relationship between questions and answers, rather than appropriate multimodal reasoning.

We propose an evaluation dataset for VCR, VCR-Visor (**VCR** benchmark dataset for **Vis**ually gr**O**unded **R**easoning), which aims to alleviate the impact of language bias for evaluating models. We reconstructed the VCR dataset so that an example of our dataset consists of two sub-questions, as shown in Figure 1. Models that rely on text-only biases will fail in one of the two sub-questions. In order to give high scores to models that correctly answer both sub-questions, we use two new metrics instead (Section 3.4).

We evaluate various VL models on our dataset. Since our dataset does not alter the distribution between questions and answers at all, we expect that models with high generalization performance on VCR also perform well on our dataset. However, all the VL models trained on VCR show much lower performance on VCR-Visor. On the other hand, the zero-shot model showed higher performance in VCR-Visor compared to VCR. We conclude that the VL models trained on VCR rely on text-based biases and VCR-Visor can measure such behaviors.

Our contributions are as follows. 1) We introduce the VCR-Visor dataset, which is designed to evaluate whether models rely on text biases for visual commonsense reasoning. 2) We provide experimental results and analysis of the models using VCR-Visor. 3) We show that methods for language bias mitigation can achieve performance improvements on VCR-Visor. However, they result in a decrease in performance on the original VCR dataset. This result indicates a significant potential for improvement in the models' generalization ability.

2 Related Works

2.1 Dealing with Biases in Visual Question Answering Benchmark Datasets

Models trained on visual question answering datasets such as VQA and VCR may achieve high performance only on in-domain test sets. Many studies have proposed out-of-distribution (OOD) datasets to measure the generalization ability of models. VQA-CP (Agrawal et al., 2018) resplits the train/test/validation set of the VQA dataset so that the answer distribution between splits is different. Models that exploit answer distribution as shortcuts fail in the VQA-CP dataset. VQA-CE (Dancette et al., 2021) and VQA-VS (Si et al., 2022) present OOD datasets targeting multimodal biases as well as unimodal biases. Ye and Kovashka (2021) proposes an OOD dataset for VCR which deals with a bias related to the referring tags.

These works target specific biases that are statistically identifiable with human supervision. On the other hand, there exist correlations that are difficult to capture explicitly between questions and answers in the VCR dataset. Our dataset addresses such correlations that can be unintended shortcuts for the task.

2.2 Visual Commonsense Reasoning

Recently, Visual commonsense reasoning has attracted a lot of attention from the community, leading to numerous benchmark datasets (Zellers et al., 2019; Park et al., 2020; Yin et al., 2021; Dong et al., 2022; Kim et al., 2022; You et al., 2022). In this study, We focus on the VCR dataset as it is the most large-scale question answering dataset and contains a wide range of scenarios.

Although the VCR dataset requires high-level reasoning and is considered to be a challenging dataset, recent models have recorded remarkable performance improvements. However, it has been found that models still have several limitations. Models can be biased towards specific regions (Yin et al., 2021) and do not fully understand low-level information in the scene (Wang et al., 2022). In this study, we examine the dependency of models on textual modality.

2.3 Datasets with Contrast Set

The differences between contrast sets can be used to estimate the model's decision boundary. Winograd Schema Challenge (Levesque et al., 2012) is composed of a pair of sentences that have minimal differences. Models have to figure out whether "**it**" refers to "trophy" or "suitcase" according to the [small/large] in the sentence "The trophy doesn't fit into the brown suitcase because **it**'s too [small/large]". VL datasets with contrast sets (Shekhar et al., 2017; Gardner et al., 2020; Hendricks and Nematzadeh, 2021; Parcalabescu et al., 2022) evaluate whether models can properly distinguish specific target elements such as objects, verbs, and locations within images when they are changed in captions. Winoground (Thrush et al., 2022) is similar to VCR-Visor in that one example is composed of two images and two captions. This dataset is designed to evaluate compositional reasoning between text and image by having each caption composed of the same words but with different structures. For example, the model must identify the image corresponding to "there is [a mug] in [some grass]" and "there is [some grass] in [a mug]". VCR-Visor is designed to evaluate whether models can find the appropriate answer based on the image, without relying on the correlation between the question and answer. To do this, each sub-question has different answers for a common question.

Reasoning Type	Words
Causal	why, how
Temporal	before, after, when, hat will, earlier, later

Table 1: Words we used to determine the reasoning type of the questions. For causal reasoning, we excluded questions that contain 'how' and 'feel' at the same time because it asks about the sentiment of the person.

3 VCR-Visor

3.1 Dataset

Our goal is to build a dataset that requires visual grounding as a necessity and penalizes models that depend on textual modality. To do this, our dataset evaluates whether the model is able to choose the correct response for each image to the same meaning query. Therefore, an example of our dataset consists of a pair of sub-questions and each sub-question has a query, image, and a pair of responses (Figure 1).

VCR-Visor deals with three types of reasoning: Causal, temporal, and explanatory reasoning. We collected causal and temporal reasoning examples from the Q $\rightarrow$ A task of the VCR dataset. We classify the reasoning type of the example according to the existence of specific words in the question (Table 1). Next, as the QA $\rightarrow$ R task of the VCR dataset is essentially an explanatory task to find an explanation for the answer, we classify the entire QA $\rightarrow$ R example as an explanatory reasoning type. In the explanatory type, we treat the combination of question and answer as a query and the rationale as the response.

3.2 Dataset Construction

We used query, image, and response pairs (Q, I, R) from the validation set of the VCR dataset. We went through a filtering process using the sentence embedding model and Natural Language Inference (NLI) model to find pairs with the same meaning query and different meaning response.

First, we calculated the embedding $E(S)$ for all of the queries and responses with Sentence-BERT model (Reimers and Gurevych, 2019). We collected a pair of examples in which the cosine similarity between queries was above the threshold and the cosine similarity between the responses was below the threshold. We used 0.9 as the threshold value. Second, We used RoBERTa (Liu et al., 2019) trained on the MNLI dataset (Williams et al., 2018) to label the relationship between the queries of the

Dataset		#Examples	#Examples per Image	Avg. Q. Len.	Avg. R. Len.	Answer-Distractor Similarity
VCR	Causal	9,848	1.59	7.99	9.46	0.452
	Temporal	1,548	1.13	8.20	8.79	0.487
	Explanatory	26,534	2.67	16.12	17.69	0.527
VCR-Visor	Causal	5,000	6.85	6.85	9.29	0.333
	Temporal	2,000	6.91	5.96	8.90	0.344
	Explanatory	3,519	3.88	11.51	15.36	0.489

Table 2: Statistisc for the VCR-Visor dataset and the VCR dataset. 'Avg. Q. Len.': Average of query length. 'Avg. R. Len.': Average of response length.

Dataset	Used Examples		Unused Examples	
	#Examples	ACC	#Examples	ACC
Causal	2,563	69.7	7,285	65.7
Temporal	607	77.3	941	69.4
Explanatory	2,055	70.7	24,479	74.1

Table 3: Comparison between examples of VCR that are used for VCR-Visor and examples that are not used. 'ACC': Accuracy of UNITER-Base trained on the VCR dataset.

collected pairs as 'entailment' or 'non-entailment'. We also labeled relationships between responses. We removed pairs where the relationship between queries is 'non-entailment' or the relationship between responses is 'entailment'. Also, pairs using the same video source were excluded.

Finally, we performed post-processing to modify the tags in some response candidates to remove the shortcuts (Ye and Kovashka, 2021). For example, if a person tag '[person2]' exists in an incorrect response and only '[person1]' exists in a query, '[person2]' in the response is modified to '[person1]'. As a result, an example of the dataset consists of two sub-questions $(Q_1, I_1, R_1, \acute{R}_2)$ and $(Q_2, I_2, \acute{R}_1, R_2)$, where $\acute{R}$ is an incorrect response and may have been post-processed.

Statistics for VCR-Visor is provided in Table 2. Since one query-response can be matched to several other query-response pairs, our dataset can create more examples than the original VCR dataset. We built more than 100,000 examples for each of the causal and temporal types, but we randomly extracted 5,000 examples and 2,000 examples for each type to avoid creating too many examples from one image. For the explanatory type, since the query consists of a question and an answer, there were relatively few query pairs with the same meaning. We made 3,519 examples and we used all of the examples.

3.3 Dataset Analysis

The average query length of the VCR-Visor dataset is shorter than that of the VCR dataset (Table 2). This is because short queries such as "Why is [person1] looking at [person2]?" are frequently used in various images, while long queries such as "Why is [person2] standing with her back turned to [person1]?" tend to be specific to the image and have few other queries with the same meaning. On the other hand, the average response length is similar in the VCR dataset and the VCR-Visor dataset.

The models show lower performance on the VCR-Visor dataset than on the VCR dataset. Hence, VCR-Visor may have cherry-picked hard examples from the VCR dataset. In order to examine whether this is true, we measure the similarity between correct responses and distractors. The more similar the Distractors are to the correct answer, the more difficult it will be to distinguish them from the correct answer. As shown in the Table 2, the similarity between correct response and distractors is lower in the VCR-Visor dataset than in the VCR dataset, which means that distractors in the VCR-Visor can be easily distinguished from the correct response. We further experimented with UNITER-Base (Chen et al., 2020) on examples from the VCR dataset that are used in VCR-Visor and examples that are not used. According to the results in Table 3, except for the explanatory type, the model performance for the examples used in VCR-Visor is higher compared to the examples that are not used. Therefore, VCR-Visor is not a dataset that uses difficult examples of the VCR dataset.

3.4 Evaluation Metrics

We use two metrics to evaluate the model on the proposed dataset. Both metrics are designed to favor a model that correctly answers both sub-questions in the example. We compute the model's prediction probability value for each correct re-

sponse to each sub-question by applying a softmax function to the model's logit value α.

$$\alpha_{i,j} = f(R_j|I_i, Q_i) \quad (1)$$

$$P_i = \frac{exp(\alpha_{i,i})}{exp(\alpha_{i,1}) + exp(\alpha_{i,2})} \quad (2)$$

Accuracy of Paired Question (APQ)
This metric tests models if they can find the correct answer for both sub-questions. This metric is very similar to the 'text score' from (Thrush et al., 2022).

$$APQ = \frac{1}{N}\sum^{N}[P_1 > 0.5 \; and \; P_2 > 0.5] \quad (3)$$

$$[x] = \begin{cases} 1 & \text{if } x = True \\ 0 & \text{if } x = False \end{cases} \quad (4)$$

Harmonic mean of Paired Question (HPQ)
This metric is designed to penalize language-dependent models. Therefore, this metric penalizes models that predict the same response as the correct answer with a high probability in two sub-questions.

$$H(x, y) = \frac{2xy}{x + y} \quad (5)$$

$$HPQ = \frac{1}{N}\sum^{N} H(P_1, P_2) \quad (6)$$

4 Experiment

4.1 Baseline Models

Random Random model selects the answer randomly.
Blind Blind model takes in the query and response as input. We use RoBERTa-Large (Liu et al., 2019) finetuned on the VCR dataset.
CLIP (Radford et al., 2021) CLIP is a dual encoder model composed of an image encoder and a text encoder. We evaluate both CLIP-Zero-shot and CLIP-Tuned which is trained on the VCR dataset. We use pretrained checkpoint of *clip-vit-base-patch16*[1].
VL-BERT (Su et al., 2020), UNITER (Chen et al., 2020), ERNIE-ViL(Yu et al., 2021) VL-BERT and UNITER are BERT-based VL models that use visual features from Faster-RCNN (Ren et al., 2015). ERNIE-ViL is an ERNIE-based (Sun et al., 2020) VL model that also utilizes visual features from Faster-RCNN. It is reported that UNITER and ERNIE-ViL show higher performance when second-stage pretraining is applied on the VCR dataset. However, for a fair comparison with other models, we used models without second-stage pretraining.
MerlotReserve (Zellers et al., 2022) MerlotReserve is pretrained using joint representation of text, image, and sound in videos. It uses Vision Transformer (Dosovitskiy et al., 2020) as the image encoder.

We provide a summary of the pretraining data, pretraining task, and vision encoder for the baseline models in Table 9.

4.2 Training and Implementation

We follow implementations from GitHub repositories for VL-BERT[2], ERNIE-ViL[3], UNITER[4], MerlotReserve [5] and CLIP[6]. We use released pretrained checkpoints for finetuning. For MerlotReserve, we use fined-tuned checkpoints. Hyperparameters for finetuning models are provided in Table 8.

Following (Teney et al., 2020), we evaluate models trained on the VCR dataset without any additional training to measure the generalization ability of models. While testing, as referring to the other sub-question in the examples can be an unintended shortcut, We input each sub-question separately into models.

4.3 Countering Language Biases

We apply RUBi (Cadene et al., 2019) and CF-VQA (Niu et al., 2021) which have been successful in reducing language bias of VQA models. We experiment these methods with CLIP and compare them with CLIP-Tuned. CLIP-Tuned is trained according to the architecture shown in Figure 3 (a), while CLIP-RUBi and CLIP-CF are trained according to the architecture shown in Figure 3 (b). CLIP-RUBi and CLIP-CF use different fusion methods and follow different causal graphs during the inference

[1] https://huggingface.co/openai/clip-vit-base-patch16

[2] https://github.com/jackroos/VL-BERT

[3] https://github.com/PaddlePaddle/ERNIE/tree/repro/ernie-vil

[4] https://github.com/ChenRocks/UNITER

[5] https://github.com/rowanz/merlot_reserve

[6] https://github.com/huggingface/transformers/tree/v4.25.1/src/transformers/models/clip

Model	Dataset	Metric	Mean	Causal	Temporal	Explanatory
Random	VCR	ACC	25.0	25.0	25.0	25.0
	VCR-Visor	APQ	25.0	25.0	25.0	25.0
Blind	VCR	ACC	64.8	58.6	58.3	77.7
	VCR-Visor	APQ	11.9	8.1	12.4	15.1
CLIP (Zero-shot)	VCR	ACC	48.3	50.5	50.6	43.9
	VCR-Visor	APQ	54.7	52.9	67.3	44.0
CLIP (Tuned)	VCR	ACC	66.9	65.3	68.3	67.1
	VCR-Visor	APQ	66.1	65.0	75.0	58.4
UNITER (Base)	VCR	ACC	71.0	66.7	72.5	73.8
	VCR-Visor	APQ	61.2	58.4	68.5	56.6
UNITER (Large)	VCR	ACC	74.2	69.8	75.3	77.6
	VCR-Visor	APQ	59.6	58.2	65.9	54.7
VL-BERT (Base)	VCR	ACC	72.8	69.4	74.2	74.9
	VCR-Visor	APQ	54.0	48.7	62.8	50.5
VL-BERT (Large)	VCR	ACC	76.5	72.0	79.2	78.2
	VCR-Visor	APQ	55.7	50.3	65.9	50.8
ERNIE-ViL (Base)	VCR	ACC	72.2	67.8	73.3	75.5
	VCR-Visor	APQ	60.9	57.4	70.7	54.6
ERNIE-ViL (Large)	VCR	ACC	77.5	73.3	79.6	79.5
	VCR-Visor	APQ	61.4	53.0	73.5	57.6
MerlotReserve (Base)	VCR	ACC	76.3	73.4	76.7	78.7
	VCR-Visor	APQ	67.4	64.5	74.2	63.4
MerlotReserve (Large)	VCR	ACC	82.6	78.9	83.9	84.9
	VCR-Visor	APQ	71.3	66.2	80.9	66.7

Table 4: Experimental results of baseline models on VCR and VCR-Visor.

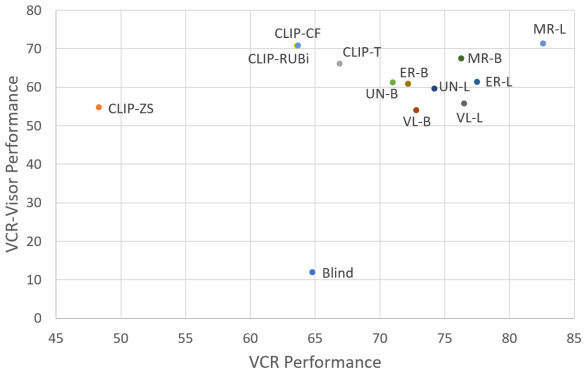
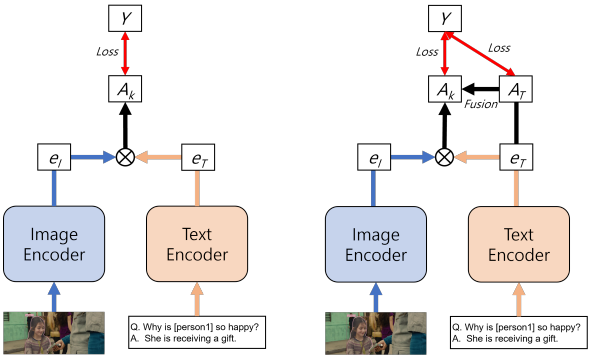


Figure 2: Performance of baseline models. The graph illustrates the results from Table 4 visually.



(a) Base Model (b) Bias-Reduced Model

Figure 3: Architecture overview of original CLIP model and bias-reduced CLIP model in the training stage. ⊗ indicates a dot product between embeddings.

process as shown in Figure 4. We implement the models based on the CF-VQA [7].

5 Results and Analysis

Although the random baseline performance of the VCR dataset and the VCR-Visor dataset is identical at 25%, they use different metrics. Additionally, as they have different structures, directly comparing their experimental results may be unfair. Therefore, we do not directly compare the performances of different models on different datasets (e.g., We do not compare the performance of UNITER on VCR with the performance of ERNIE-ViL on VCR-Visor).

[7] https://github.com/yuleiniu/cfvqa

We provide main results in Table 4. Figure 2 visually presents the same results. We have added the performance on the VCR dataset. We show the results of the APQ metric only for simplicity, and results including the HPQ metric are provided in Table 10.

5.1 Penalizing language dependent models

We designed the VCR-Visor dataset to ensure that models that rely on language fail. Blind model, which is totally dependent on language, exhibits much lower performance than the random baseline on VCR-Visor. Thus, we conclude that the VCR-Visor dataset is well-constructed for its intended

Model	VCR				VCR-Visor			
	Mean	Causal	Temporal	Explanatory	Mean	Causal	Temporal	Explanatory
CLIP-Tuned	66.9	65.3	68.3	67.1	66.1	65.0	75.0	58.4
CLIP-RUBi	63.6	64.8	66.3	59.8	70.7	70.0	79.4	62.6
CLIP-CF	63.7	65.1	65.4	60.7	70.8	69.3	80.5	62.7

Table 5: Results of language bias mitigation methods.

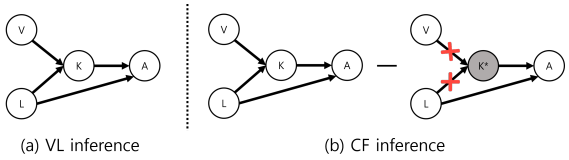


Figure 4: Causal graph of VL models. CLIP-Tuned and CLIP-RUBi models have the causal graph structure as (a) and CLIP-CF model has (b) in the inference stage. The causal graph used by us is a simplified version of the CF-VQA (Niu et al., 2021), and does not consider the V→A branch.

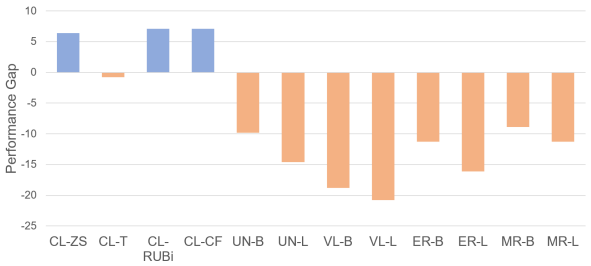


Figure 5: Performance gap of VL models between the VCR-Visor and original VCR. CL-ZS: CLIP-Zero-shot. CL-T: CLIP-Tuned. L: Large. B: Base. UN: UNITER. VL: VL-BERT. ER: ERNIE-ViL. MR: MerlotReserve.

purpose. On the other hand, in an ideal case, the performance of Blind model should be zero, but the performance is slightly higher than zero. This is because Blind model does not predict identical scores to query-response pairs with the same meaning, as shown in Figure 1. Additionally, we note that since the dataset is generated in an automated manner including post-processing, there can be instances that contain shortcuts for the answer.

5.2 VCR vs. VCR-Visor

High performance on VCR does not guarantee high performance on VCR-Visor. The Pearson correlation coefficient value between the models' performance on VCR and VCR-Visor is 0.483, indicating a linear correlation, however, there are many exceptional cases. As shown in Table 4, CLIP-Zero-shot has the lowest performance among the baseline models on VCR, but it shows a higher performance than VL-BERT-Base on VCR-Visor. Especially, UNITER-Large has lower performance than UNITER-Base on VCR-Visor.

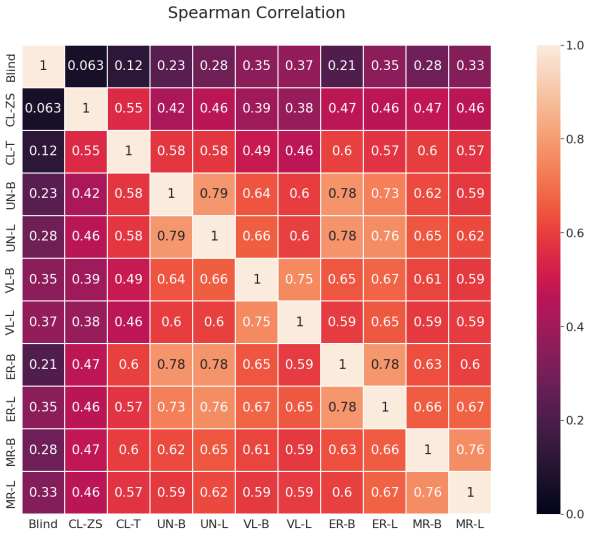


Figure 6: Spearman correlations between predictions of models on causal reasoning examples in the VCR-Visor dataset.

Figure 5 illustrates the performance gap between VL models on VCR-Visor and VCR. CLIP-Zero-shot exhibits higher performance on VCR-Visor compared to VCR. Conversely, CLIP-Tuned demonstrates slightly lower performance on VCR-Visor compared to VCR. Therefore, the zero-shot model exhibits low language dependency, whereas finetuning causes language dependency of the model.

We also discover that the performance gap is larger in large models compared to base models. As shown in Figure 5, for UNITER, VL-BERT, ERNIE-ViL, and MerlotReserve models, the performance gap is greater in large models. This implies that language-based reasoning has contributed more to the performance improvement of large models than multimodal reasoning.

One of the most important questions is how we can determine which model has better performance. We believe that VCR and VCR-Visor should be employed as complementary evaluation sets. For example, We can compare models

with similar performance on VCR in Figure 2. VL-BERT-Large shows slightly higher VCR performance than Merlot-Reserve-Base, but Merlot-Reserve-Base shows much higher VCR-Visor performance. Therefore, we conclude that Merlot-Reserve-Base shows much better performance than VL-BERT-Large.

5.3 Correlation between models' predictions

We measured the spearman correlation between the predictions of models in VCR-Visor (Figure 6). The correlation between the CLIP Zero-shot and Blind model is almost zero, indicating that these models make judgments based on entirely different evidence. This explains the reason why these models are situated in entirely different positions in Figure 2, implying that VCR and VCR-Visor evaluate the models from different perspectives in a complementary manner. Additionally, in accordance with the analysis in Section 5.2, large models have a higher correlation with blind models compared to base models.

5.4 Mitigating language dependence

Table 5 shows that language bias mitigation methods successfully alleviate the model's language dependence. CLIP-RUBi and CLIP-CF show improved performance in all reasoning types compared to CLIP-Tuned model. At the same time, these models show significantly lower performance on the original VCR compared to CLIP-Tuned. The result of performance deterioration in the original dataset is very similar to the experimental results in the VQA dataset (Cadene et al., 2019; Niu et al., 2021). We conjecture that while these methods noticeably reduce the model's dependence on language modality, they may impair some beneficial language inference capabilities.

5.5 Fusion methods

We compare early fusion models (VL-BERT, UNITER, ERNIE-ViL, MerlotReserve) and late fusion models (CLIP models). While late fusion models use simple methods such as linear projection and cosine similarity for interactions between modalities, early fusion models employ cross-modal self-attention mechanisms. As shown in Figure 5, late fusion models, except for CLIP-Tuned, demonstrate higher performance on VCR-Visor than on VCR. On the other hand, early fusion models demonstrate much lower performance on VCR-Visor than on VCR.

From these experimental results, we think that early fusion models tend to rely more on language, compared to late fusion models. We believe this is because late fusion models make predictions by computing the cosine similarity between modalities, which means that they cannot give higher weight to one of the modalities than the other. On the other hand, as demonstrated in (Cao et al., 2020), early fusion models can assign lower weights to visual tokens through self-attention.

5.6 Visual features

On VCR-Visor, large models of VL-BERT and ERNIE-ViL show relatively small performance improvements of 1.7 and 0.5 respectively compared to their base models. UNITER even shows a decrease in performance of 1.6. In contrast, MerlotReserve shows a relatively large performance improvement of 3.9 in the large model. We conjecture that the reason for these results is the limitations of the pre-extracted Faster-RCNN visual representations used by VL-BERT, ERNIE-ViL, and UNITER. Meanwhile, MerlotReserve-Large uses a larger visual feature extractor than MerlotReserve-Base. The observation that visual features act as a limiting factor is in line with the findings in previous studies (Zhang et al., 2021; Salin et al., 2022).

6 Conclusion

We propose the VCR-Visor dataset, designed to quantitatively measure the extent to which models depend on the text modality in visual commonsense reasoning. Models trained on the original VCR dataset generally exhibit a significant drop in performance when evaluated on VCR-Visor, suggesting that their success on VCR may largely stem from reliance on language cues. Our results show that increasing model size can amplify language dependence, highlighting the importance of improving visual representations. Furthermore, we find that language bias mitigation techniques can greatly enhance performance on VCR-Visor; however, these techniques often lead to a decrease in performance on the original VCR dataset. In summary, VCR-Visor provides a complementary benchmark to VCR, focusing on evaluating models' ability to reason without being biased toward a particular modality. We hope that our dataset and comprehensive analysis will foster future research in this area.

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.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Remi Cadene, Corentin Dancette, Hedi Ben younes, Matthieu Cord, and Devi Parikh. 2019. Rubi: Reducing unimodal biases for visual question answering. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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 *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI*, page 565–580, Berlin, Heidelberg. Springer-Verlag.

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 *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, page 104–120, Berlin, Heidelberg. Springer-Verlag.

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.

Qingxiu Dong, Ziwei Qin, Heming Xia, Tian Feng, Shoujie Tong, Haoran Meng, Lin Xu, Zhongyu Wei, Weidong Zhan, Baobao Chang, Sujian Li, Tianyu Liu, and Zhifang Sui. 2022. Premise-based multimodal reasoning: Conditional inference on joint textual and visual clues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 932–946, Dublin, Ireland. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. 2022. A survey of vision-language pre-trained models. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5436–5443. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

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.

Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Lisa Anne Hendricks and Aida Nematzadeh. 2021. Probing image-language transformers for verb understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3635–3644, Online. Association for Computational Linguistics.

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.

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.

Hyounghun Kim, Abhay Zala, and Mohit Bansal. 2022. CoSIm: Commonsense reasoning for counterfactual scene imagination. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 911–923, Seattle, United States. Association for Computational Linguistics.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vision*, 123(1):32–73.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2022. VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8253–8280, Dublin, Ireland. Association for Computational Linguistics.

Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. 2020. Visualcomet: Reasoning about the dynamic context of a still image. In *European Conference on Computer Vision*, pages 508–524. Springer.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Emmanuelle Salin, Badreddine Farah, Stéphane Ayache, and Benoit Favre. 2022. Are vision-language transformers learning multimodal representations? a probing perspective. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11248–11257.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017. FOIL it! find one mismatch between image and language caption. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 255–265, Vancouver, Canada. Association for Computational Linguistics.

Qingyi Si, Fandong Meng, Mingyu Zheng, Zheng Lin, Yuanxin Liu, Peng Fu, Yanan Cao, Weiping Wang, and Jie Zhou. 2022. Language prior is not the only shortcut: A benchmark for shortcut learning in vqa. *ArXiv*, abs/2210.04692.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*.

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. 2017. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223, Vancouver, Canada. Association for Computational Linguistics.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8968–8975.

Damien Teney, Ehsan Abbasnejad, Kushal Kafle, Robik Shrestha, Christopher Kanan, and Anton van den Hengel. 2020. On the value of out-of-distribution testing: An example of goodhart's law. In *Advances in Neural Information Processing Systems*, volume 33, pages 407–417. Curran Associates, Inc.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5238–5248.

Zhecan Wang, Haoxuan You, Yicheng He, Wenhao Li, Kai-Wei Chang, and Shih-Fu Chang. 2022. Understanding me? multimodal evaluation for fine-grained visual commonsense. In *EMNLP*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

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.

Da Yin, Liunian Harold Li, Ziniu Hu, Nanyun Peng, and Kai-Wei Chang. 2021. Broaden the vision: Geo-diverse visual commonsense reasoning. In *EMNLP*.

Haoxuan You, Rui Sun, Zhecan Wang, Kai-Wei Chang, and Shih-Fu Chang. 2022. Find someone who: Visual commonsense understanding in human-centric grounding. In *EMNLP-Finding*.

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3208–3216.

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, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. 2022. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16375–16387.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Making visual representations matter in vision-language models. *CVPR 2021*.

Wangchunshu Zhou, Yan Zeng, Shizhe Diao, and Xinsong Zhang. 2022. VLUE: A multi-task multi-dimension benchmark for evaluating vision-language pre-training. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 27395–27411. PMLR.

A Language shortcuts in the VCR dataset

To investigate the extent of language shortcuts present in the VCR dataset, we conducted pilot experiments. We categorized examples in the dataset into two levels of language shortcuts: 1) Word-level language shortcut, and 2) Sentence-level language shortcut. An example with a word-level language shortcut is one where the correct answer can be identified solely based on the presence of certain words in the question and the choices. For instance, if the question includes words like 'happy' or 'she', the correct answer is likely to contain words such as 'smile' or 'laugh'. To isolate only the word-level information, we created modified VCR data by randomly shuffling the word order in both the question and the answer sentence. An example with a sentence-level language shortcut is one where the correct answer can be determined using only the question and the choices.

We measured each language shortcut through experiments using language models. The language model receives the question and one of the answer candidates as input and assigns a score. Among the four candidates, the one with the highest score is selected as the answer. To minimize the influence of model-specific characteristics on shortcut detection, we used two language models: BERT and RoBERTa. Additionally, to rule out cases where the model might guess correctly by chance, each language model was trained with three different random seeds. An example was classified as 'Biased' if all six models consistently selected the correct answer, as 'Non-Biased' if none of them selected the correct answer, and as 'Neutral' otherwise. The results are presented in Table 6.

A.1 Word-level shortcut v.s. Sentence-level shortcut

The VCR dataset exhibits stronger sentence-level shortcuts than word-level shortcuts. Therefore, understanding sentence-level semantics is more important than identifying superficial word-level correlations between the question and the answer.

A.2 Question Answering task (Q → A) v.s. Justification task (QA→R)

Word-level shortcuts are more prominent in the Q→A task, while sentence-level shortcuts are more prominent in the QA→R task. Thus, the QA→R task contains fewer superficial word-level correlations and requires greater sentence comprehension

	Biased (%)	Non-Biased (%)	Neutral (%)	Total (%)
Word-level (Q->A)	39.2	37.0	23.8	100
Word-level (QA->R)	29.0	48.1	22.9	100
Sentence-level (Q->A)	50.8	27.4	21.8	100
Sentence-level (QA->R)	63.6	14.8	21.6	100

Table 6: Results of text-based shortcut detecting experiment in the VCR dataset.

Dataset	#Examples
PMR	1,538
PMR-P	3,502
COSIM	800
COSIM-Q	154
COSIM-C	182

Table 7

ability.

On the other hand, although the Q→A task exhibits a higher degree of superficial correlation, there are fewer examples that can be answered correctly using text alone compared to the QA→R task. Therefore, visual understanding is more crucial for achieving high performance in the Q→A task than in the QA→R task.

B Model implementations

We concatenated the question and answer for the Q → A task, and concatenated the question, answer, and rationale for the QA → R task to input into the model. Tags such as '[person1]' were converted to ''[person1]'', and tags like '[object1]' were replaced with the corresponding object names (e.g., 'bag') before being input to the model. The hyperparameters for model training are provided in Table 8. We evaluated the model using the checkpoint that achieved the highest performance on the VCR validation set.

C Applicability of Our Dataset Construction Methods

In this section, we verify the applicability of our dataset construction method to various visual commonsense reasoning datasets and video question answering datasets. We used the PMR (Dong et al., 2022) and COSIM (Kim et al., 2022) datasets.

The PMR dataset (Dong et al., 2022) involves inferring a person's next action given a premise such as the person's psychological state in an image. Using the PMR dataset, we created a new PMR-P dataset by pairing questions with semantically equivalent premises. The COSIM dataset (Kim et al., 2022) requires counterfactual reasoning about scenarios depicted in images. Each COSIM example consists of a question, a counterfactual change, and an answer. We constructed two new datasets from COSIM: COSIM-Q, which pairs semantically equivalent questions, and COSIM-C, which pairs semantically equivalent counterfactual changes.

The dataset construction results are shown in Table 7. For the COSIM dataset, which has only 800 validation examples, we were able to construct 154 examples for COSIM-Q and 182 examples for COSIM-C. In contrast, the PMR dataset contains 1,538 examples, allowing us to construct 3,502 new examples. These results demonstrate that our dataset construction method can be applied to visual-language question answering tasks, provided that a sufficient number of examples are available.

D Correlation between model predictions

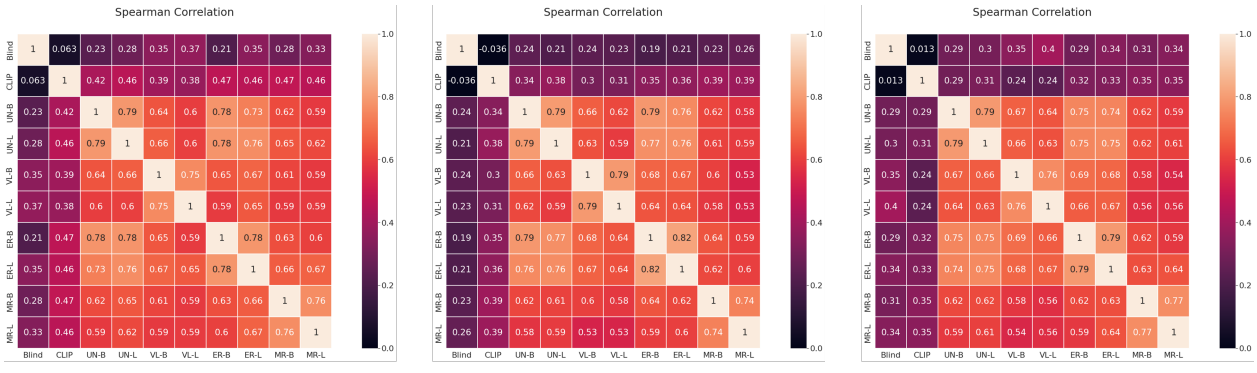
In addition to the model prediction correlations for VCR-Visor causal reasoning presented in the main text, we provide the model prediction correlations for temporal reasoning and explanatory reasoning as well (Figure 7).

Model	Optimizer	Batch Size	Learning Rate	#Training Steps/Epochs
RoBERTa (Large)	AdamW	{16}	{2e-5}	{3, 5}
VL-BERT (Base)	SGD	{16}	{7e-5}	{10, 20}
VL-BERT (Large)	SGD	{16}	{7e-5}	{10, 20}
UNITER (Base)	AdamW	{64}	{3e-5, 6e-5}	{32000, 64000}
UNITER (Large)	AdamW	{64}	{3e-5, 6e-5}	{32000, 64000}
ERNIE-ViL (Base)	Adam	{16}	{2e-5}	{26640, 53280}
ERNIE-ViL (Large)	Adam	{16, 64}	{2e-5}	{26640, 53280}
CLIP-Tuned	AdamW	{16, 32}	{1e-6, 1e-7}	{5}
CLIP-RUBi	AdamW	{16, 32}	{1e-6, 1e-7}	{5}
CLIP-CF	AdamW	{16, 32}	{1e-6, 1e-7}	{5}

Table 8: Hyperparameters for model training. We used the finetuned model checkpoint of MerlotReserve provided at https://github.com/rowanz/merlot_reserve.

Model	Pretraining Data	Pretraining Task	Vision Encodeer
VL-BERT	CC	MLM, MRC	Faster-RCNN
UNITER	CC, COCO, VG, SBU	MLM, MRC, ITM, WRA, MRFR	Faster-RCNN
ERNIE-ViL	CC, SBU	MLM, MRP, ITM	Faster-RCNN
MerlotReserve	YT-Temporal-1B (Self-collected from YouTube)	CST	ViT
CLIP	400M Image-Text pairs (Self-collected from the Internet)	CL	ViT, ResNet

Table 9: Summary of VL models. CC: Conceptual Captions (Sharma et al., 2018). COCO: Microsoft COCO (Lin et al., 2014), VG: Visual Genome (Krishna et al., 2017) . SBU: SBU captions (Ordonez et al., 2011). MLM: Masked Language Modeling. MRC: Masked Region Classification. ITM: Image Text Matching. WRA: Word Region Alignment. MRFR: Masked Region Feature Regression. MRP: Masked Region Prediction. CST: Contrastive Span Training (Zellers et al., 2022). CL: Contrastive Learning.



(a) Causal reasoning (a) Temporal reasoning (a) Explanatory reasoning

Figure 7: Correlation between model predictions.

Model	Dataset	Metric	Mean	Causal	Temporal	Explanatory
Random	VCR	ACC	25.0	25.0	25.0	25.0
	VCR-Visor	APQ	25.0	25.0	25.0	25.0
		HPQ	50.0	50.0	50.0	50.0
Blind	VCR	ACC	64.8	58.6	58.3	77.7
	VCR-Visor	APQ	11.9	8.1	12.4	15.1
		HPQ	35.0	28.6	40.5	36.0
CLIP (Zero-shot)	VCR	ACC	48.3	50.5	50.6	43.9
	VCR-Visor	APQ	54.7	52.9	67.3	44.0
		HPQ	62.7	61.7	70.7	55.7
CLIP (Tuned)	VCR	ACC	66.9	65.3	68.3	67.1
	VCR-Visor	APQ	66.1	65.0	75.0	58.4
		HPQ	73.7	72.9	80.9	67.3
CLIP (RUBi)	VCR	ACC	63.6	64.8	66.3	59.8
	VCR-Visor	APQ	70.7	**70.0**	79.4	62.6
		HPQ	**74.4**	**73.5**	81.4	68.3
CLIP (CF)	VCR	ACC	63.7	65.1	65.4	60.7
	VCR-Visor	APQ	70.8	69.3	80.5	62.7
		HPQ	**74.4**	73.1	81.6	68.5
UNITER (Base)	VCR	ACC	71.0	66.7	72.5	73.8
	VCR-Visor	APQ	61.2	58.4	68.5	56.6
		HPQ	67.1	64.8	72.6	63.8
UNITER (Large)	VCR	ACC	74.2	69.8	75.3	77.6
	VCR-Visor	APQ	59.6	58.2	65.9	54.7
		HPQ	66.4	62.8	72.6	63.8
VL-BERT (Base)	VCR	ACC	72.8	69.4	74.2	74.9
	VCR-Visor	APQ	54.0	48.7	62.8	50.5
		HPQ	58.5	53.4	65.9	56.3
VL-BERT (Large)	VCR	ACC	76.5	72.0	79.2	78.2
	VCR-Visor	APQ	55.7	50.3	65.9	50.8
		HPQ	58.9	53.6	68.1	55.1
ERNIE-ViL (Base)	VCR	ACC	72.2	67.8	73.3	75.5
	VCR-Visor	APQ	60.9	57.4	70.7	54.6
		HPQ	66.7	64.5	74.3	61.4
ERNIE-ViL (Large)	VCR	ACC	77.5	73.3	79.6	79.5
	VCR-Visor	APQ	61.4	53.0	73.5	57.6
		HPQ	66.4	59.1	76.2	63.9
MerlotReserve (Base)	VCR	ACC	76.3	73.4	76.7	78.7
	VCR-Visor	APQ	67.4	64.5	74.2	63.4
		HPQ	71.1	68.7	76.8	67.9
MerlotReserve (Large)	VCR	ACC	**82.6**	**78.9**	**83.9**	**84.9**
	VCR-Visor	APQ	**71.3**	66.2	**80.9**	**66.7**
		HPQ	74.2	69.8	**82.8**	**70.0**

Table 10: Experimental results of all baseline models. The best performance for each metric is highlighted in **bold**.