Appendices

A. Qualitative Examples

Two sets of figures are given to show qualitative examples of our datasets and models. Figure 5 shows the explanations generated by the models (and the ground-truth) for two images of each dataset. We also display the e-ViL S_E score of each generated explanation, which was obtained through our human evaluation framework. In some of the images, such as in Figures 5b and 5f, we can see that e-UG provides better, more image-grounded explanations.

In Figure 6 we again show two images per dataset. These examples illustrate the key differences between the different datasets. VCR has many questions that require substantial commonsense reasoning and general knowledge. For example, to explain the answer to the question in Figure 6e, one needs to know that person 1 is wearing a T-shirt of the classic rock band Guns N' Roses. For VQA-X (Figures 6c and 6d), we show two examples where a generic explanation, that is not necessarily grounded in the image, will often suffice (this is a general limitation of this dataset). The explanation "Because there is a person on a surfboard" and "Because there is a bed in the room" will, in most cases, be correct with respect to the question and answer, regardless of the image. The examples for e-SNLI-VE in Figures 6a and 6b both require the explanations to describe image-specific characteristics in order to be meaningful. In Figure 6a, a valid explanation would have to pick out a concrete element from the image to explain why it is a contradiction.

B. e-SNLI-VE

This section contains a datasheet on e-SNLI-VE, as well as further information on its pre-processing. Details on the filters are given in Section B.4. Details on the MTurk evaluation can be found in Appendix D.

B.1. e-SNLI-VE Datasheet

The questions in this section will be answered predominantly with respect to the changes that were applied on top of (e-)SNLI, SNLI-VE, and Flickr30k. We use the datasheet form from Gebru et al. [20].

B.1.1 Motivation

For what purpose was the dataset created? The dataset was created for the purpose of extending the range of existing VL-NLE datasets with a large-scale dataset that requires fine-grained reasoning.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)? The dataset was created by researchers from the University of Oxford. It builds on existing datasets which involved other institutions (NEC Laboratories America for SNLI-VE) and universities (Stanford University for SNLI, University of Illinois at Urbana-Champaign for Flickr30k, University of Oxford for e-SNLI).

B.1.2 Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Photos (some with people) and natural language sentences.

How many instances are there in total? In total, there are 430,796 instances.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? The dataset contains a reduced sample of the original 570k sentence pairs from SNLI [10]. It has been reduced because various filtering methods were applied to remove noise that occurred from combining e-SNLI and SNLI-VE. The filtering steps disproportionately affect the "neutral" class.

What data does each instance consist of? Each instance consists of an image, a natural language hypothesis, a label that classifies the image-hypothesis pair as entailment, contradiction, or neutral, and a natural language explanation that explains why the label was given.

Is any information missing from individual instances? No, all instances contain the complete the information described above.

Are relationships between individual instances made explicit? Yes. Some instances refer to the same image, which is indicated via their image ID.

Are there recommended data splits? Yes, the train, dev, and test splits are given with the release of the dataset.

Are there any errors, sources of noise, or redundancies in the dataset? The labels and explanations were originally annotated for textual premise-hypothesis pairs. By replacing the textual premise with an image, noise occurs. Despite our best efforts to filter out this noise, a considerable error rate remains.

Is the dataset self-contained, or does it link to or otherwise rely on external resources? The dataset needs to be linked with Flickr30k images, which are publicly available.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? No.

	Human Evaluation:
Hypothesis: A dog is playing with a cat.	
Relation: Contradiction	
GT Explanation: A man running and a dog playing with a cat are two very distinct activities.	
PJ-X: a dog is not a cat	0.00
FME: a dog is not a cat	0.17
RVT: A cat is not a dog.	0.00
e-UG: A dog is not a football player.	0.56

(a) e-SNLI-VE.

	Human Evaluation:
Hypothesis: The lady is the owner of the store.	
Relation: Neutral	
GT Explanation: We cannot tell from this picture if the lady is the owner of the store.	
PJ-X: a woman looking at a microscope does not imply that she is looking for the store	0.56
FME: a woman can be a man or a woman	0.17
RVT: Just because a lady is holding a book does not mean she is the owner of the store.	0.67
e-UG: Just because a lady is working at a store does not mean she is the owner.	1

(b) e-SNLI-VE.

	Human Evaluation:
Question: What sport are they playing?	
Answer: tennis	
GT Explanation: they are holding tennis rackets inside the court.	
PJ-X: the man is holding a tennis racket	1
FME: they are holding tennis rackets	0.89
RVT: the man is holding a racket	0.89
e-UG: they are holding tennis rackets	1

(c) VQA-X.

	Human Evaluation:
Question: What game is the guy playing?	
Answer: frisbee	
GT Explanation: he is about to catch one in his hand.	
PJ-X: he is holding a frisbee	0.22
FME: the man is holding a frisbee	0.22
RVT: he is holding a disc	0.33
e-UG: he is throwing a frisbee.	0.67

(d) VQA-X.

	Human Evaluation:
Question: What is 1 doing ?	
Answer: 1 is confronting 3.	
GT Explanation: 3 looks uncomfortable and 1 looks as if he is not happy.	
PJ-X: 1 is looking at 3 with a smile on his face	0.00
FME: 1 is looking at 3 with a frown	0.78
RVT: 1 is looking at 3 with a look of anger on his face.	0.5
e-UG: 1 is facing 3 and is speaking to him.	0.89

(e) VCR.

	Human Evaluation:
Question: What is in wineglass ?	
Answer: There ' s champagne in wineglass.	
GT Explanation: There 's a bottle of champagne near it.	
PJ-X: there is a plate of food in front of her and she is holding a fork	0.33
FME: wineglass is a wine glass and wineglass is a wine glass	0.44
RVT: wineglass is a wineglass and wineglass is a wineglass.	0.67
e-UG: The glass is full of champagne and there is a champagne bottle in front of it.	1

(f) VCR.

Figure 5: Pair of examples from the test set of each dataset. We display the ground-truth (GT) explanation, as well as the generated explanations of each model and their predicted human evaluation score S_E.

Hypothesis: The man rode his bike. **Relation:** Contradiction

(a) e-SNLI-VE.

Hypothesis: A male has a hat on. **Relation:** Entailment

(b) e-SNLI-VE.

Question: What sport is this person doing? **Answer:** surfing

(c) VQA-X.

Question: What room is this? **Answer:** bedroom

(d) VQA-X.

Question: What kind of music does 1 like?
Answer: He likes classic rock.

(e) VCR.

Question: What is 2 doing?
Answer: 2 is communicating with someone outside of the room to give them instruction.

(f) VCR.

Figure 6: Representative examples from each dataset.

B.1.3 Collection Process

How was the data associated with each instance acquired? Hypothesises and explanations were annotated by people. SNLI-VE combined e-SNLI and Flickr30k by replacing the textual premise by an image. This was possible because the textual premises in SNLI are all captions of Flickr30k images. e-SNLI-VE was obtained by associating the explanations from SNLI with SNLI-VE. We used MTurk to reannotate the labels and explanations for the neutral class in the validation and test set. Numerous validation steps have been used to measure the effectiveness of merging, re-annotating, and filtering the dataset.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? Software program and manual human curation.

B.1.4 Preprocessing/Cleaning/Labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? Various filters were used to remove noise. We used a false neutral detector (details in Section 3.1), a keyword filter (details in Section 3.2), a similarity filter (details in Section 3.2), and an uncertainty filter (details in Section 3.2). We also reannotated all neutral examples in the validation and test set.

B.1.5 Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? The dataset is publicly released and free to access.

B.1.6 Maintenance

Who is supporting/hosting/maintaining the dataset? The first author of this paper.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)? The first author of this paper can be contacted via the email address given on the title page.

B.2. Relabeling e-SNLI-VE via MTurk

In this work, we collect new labels and explanations for the neutral pairs of the validation and test sets of e-SNLI-VE. We provide workers with the definitions of entailment, neutral, and contradiction for image-sentence pairs and one example for each label. As shown in Figure 7, for each image-sentence pair, workers are required to (a) choose a label, (b) highlight words in the sentence that led to their label decision, and (c) explain their decision in a comprehensive and concise manner, using at least half of the words that they highlighted. We point out that it is likely that requiring an explanation at the same time as requiring a label has a positive effect on the correctness of the label, since having to justify in writing the picked label may make annotators pay an increased attention. Moreover, we implemented additional quality control measures for crowdsourced annotations, such as (a) collecting three annotations for every input, (b) injecting trusted annotations, and (c) restricting to annotators with at least 90% previous approval rate.

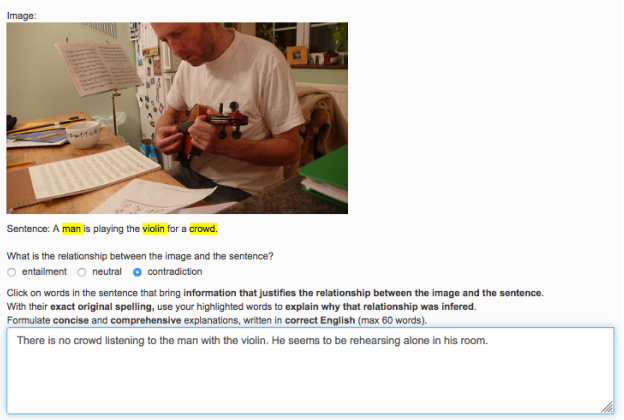


Figure 7: A snapshot of the annotation interface that was used to manually reannotate the neutral labels in the validation and test sets of e-SNLI-VE.

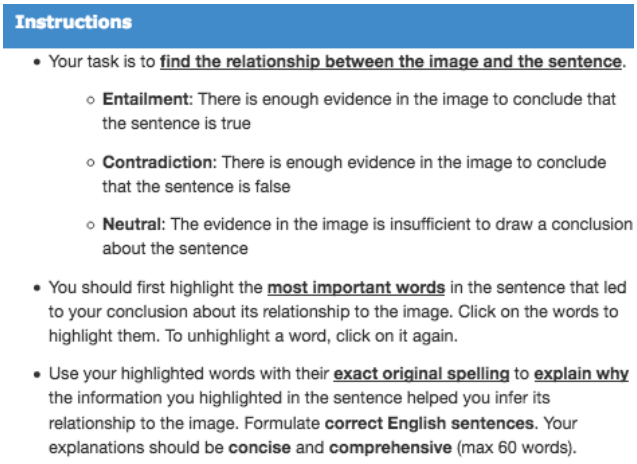


Figure 8: A snapshot of the instructions that were provided to the workers that reannotated the neutral labels in the validation and test sets of e-SNLI-VE.

There were 2,060 workers in the annotation effort, with an average of 1.98 assignments per worker and a standard deviation of 5.54. No restriction was put on the workers' location. Each assignment consisted of a set of 10 image-sentence pairs. The instructions are shown in Figure 8. The annotators were also guided by three examples, one for each label. For each assignment of 10 questions, one trusted annotation with known label was inserted at a random position, as a measure to control the quality of label annotation. Each assignment was completed by three different workers.

To check the success of our crowdsourcing, we manually assessed the relevance of explanations among a random subset of 100 examples. A marking scale between 0 and 1 was used, assigning a score of k/n when k required attributes were given in an explanation out of n. We report an 83.5% relevance of explanations from workers.

B.3. Ambiguity in e-SNLI-VE

We noticed that some instances in SNLI-VE are ambiguous. We show some examples with justifications in Figures 10, 9, and 11. In order to have a better sense of this ambiguity, three authors of this paper independently annotated 100 random examples. All three authors agreed on 54% of the examples, exactly two authors agreed on 45%, and there was only one example on which all three authors disagreed. We identified the following three major sources of ambiguity: (1) mapping an emotion in the hypothesis to a facial expression in the image premise, e.g., "people enjoy talking", "angry people", "sad woman". Even when the face is seen, it may be subjective to infer an emotion from a static image, (2) personal taste, e.g., "the sign is ugly", and (3) lack of consensus on terms such as "many people" or "crowded".

In our crowdsourced re-annotation effort, we accounted for this by removing an instance if all three annotator disagreed on the label (5.2% for validation and 5.5% test set). Otherwise we choose the majority label. Looking at the 18 instances where we disagreed with the label assigned by MTurk workers, we noticed that 12 were due to ambiguity in the examples, and 6 were due to workers' errors.

B.4. Details on Filters

In Table 7, we provide a quantitative analysis of the effects our filters had on the dataset. The accuracies are obtained from our hand-annotated subset of 535 examples. On this subset, we first annotated every image-sentence pair as Entailment, Neutral, or Contradiction. Accuracies are obtained by comparing our own annotation with the dataset annotation. Note that we obtain higher error rates for the Entailment and Contradiction classes (9.7% and 8.6%) than what the authors of the original paper found [48] (less than 1%). One explanation for that could be the ambiguity that is inherent in the task. The share of bad explanations is obtained by evaluating every explanation as *bad*, *okay*, or *great*. If the label is wrong, the explanation is automatically deemed *bad*, as it will try to explain a wrong answer.

Note that in e-SNLI, the authors have found that the human annotated explanations have an error rate of 9.6%

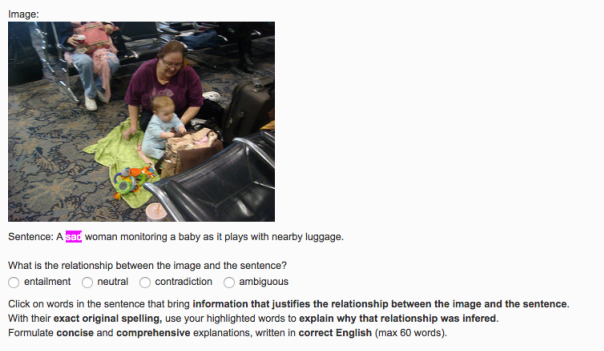


Figure 9: Ambiguous SNLI-VE instance. Some may argue that the woman's face betrays sadness, but the image is not quite clear. Secondly, even with better resolution, facial expression may not be a strong enough evidence to support the hypothesis about the woman's emotional state.

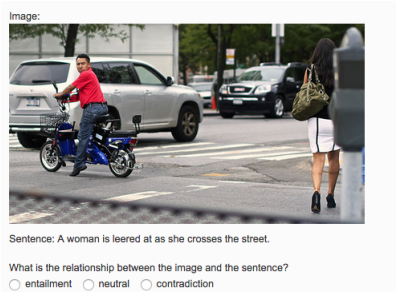


Figure 10: Ambiguous SNLI-VE instance. The lack of consensus is on whether the man is "leering" at the woman. While it is likely the case, this interpretation in favour of entailment is subjective, and a cautious annotator would prefer to label the instance as neutral.

Premise	Hypothesis	Label
	Kindergarten students napping	Neutral

Figure 11: Ambiguous SNLI-VE instance. Some may argue that it is impossible to certify from the image that the children are kindergarten students, and label the instance as neutral. On the other hand, the furniture may be considered as typical of kindergarten, which would be sufficient evidence for entailment.

(19.6% on entailment, 7.3% on neutral, 9.4% on contradiction), which serves as an upper bound of what could be achieved in terms of dataset cleaning.

An illustrative example for the motivation of the false neutral detector is given in the main paper in Figure 2. Examples for the keyword and similarity filters are given in Figures 12 and 13, respectively.

Textual premise: Older man sits and plays the accordion while young girl watches.
Hypothesis: An old man plays an instrument while a young child watches.
Label: Entailment
Explanation: An accordion is a type of instrument, also "child" is a synonym for "young girl" and "old man" is a rephrasing of "older man".

Figure 12: The use of the words "synonym" and "rephrasing" makes it clear that the explanation is overly focused on the linguistic features of the textual premise.

Textual Premise: A mother stands in a kitchen holding a small baby
Hypothesis: A mother is holding a small baby.
Label: Entailment
Explanation: A mother standing in the kitchen holding a small baby is the same as a mother holding a small baby.

Figure 13: The textual premise and hypothesis are almost identical sentences, which led to a low-quality explanation.

C. Benchmark Models

This section contains further details on the models that are compared in this benchmark.

C.1. Model Architectures

PJ-X. The PJ-X model [37] provides multimodal explanations for VQA tasks and was originally evaluated on VQA-X. Its M_T module consists of a simplified MCB network [18] that was pre-trained on VQA v2.

We implemented PJ-X in PyTorch following closely the authors' implementation in Caffe[5]. To address numerical optimization problems, we replaced the L2 normalization in the decoder with LayerNorm [7], as the original normalization zeroed gradients for earlier model parts. Additionally, we added gradient clipping of 0.1 to prevent too large gradients. To adapt PJ-X for multiple-choice question-answering in VCR, we follow the approach in the original VCR paper [50].

FME. The model introduced by Wu and Mooney [46], which we will refer to as FME (Faithful Multimodel Explanations), puts emphasis on producing faithful explanations. In particular, it aims to ensure that the explanation utilizes the same visual features that were used to produce the answer. Their code is not publicly available and we, therefore, re-implemented their base model according to the instructions in the paper. We chose the base model, as it was trained on the entire VQA-X 29.5K train split and the modifications of the other variations were difficult to re-implement from the descriptions in the paper. Our re-implementation of FME is based on a frozen modified UpDown [3] VQAv2 pre-trained VL-model.

Similarly to PJ-X, we also train FME with a gradient clipping of 0.1. To adapt FME for multiple-choice QA in VCR, we follow the approach in the original VCR paper [50].

RVT. The Rationale-VT Transformer (RVT) model [34] uses varying vision algorithms to extract information from an image and then feeds this information, the ground-truth answer, and the question to a pre-trained GPT-2 language model [38], which yields an explanation. As they omit the question answering part, we extend their model by an answer prediction module to allow for a fair comparison and to get a sense of the overall performance. We use their overall most effective visual input[6], which are the tags of the objects detected in the image. As task model M_T, we use BERT [16], which takes as input the object tags and the question, and predicts the answer.

C.2. Joint or Separate Training.

All the VL-NLE models M in this work consist of M_T and M_E modules, which can either be trained jointly or separately. For the RVT model, training jointly would make no difference, as the explanation generation is not conditioned on a learnable representation in M_T (but instead on the fixed object tags for each image). For all other models,

[5] https://github.com/Seth-Park/MultimodalExplanations

[6] It obtained the highest visual plausibility score averaged across all datasets.

	Dataset Size			Share of wrong labels				Share of bad explanations			
	Train Set	Val Set	Test Set	All	E	N	C	All	E	N	C
Raw	529,505	17,554	17,899	19.3%	9.7%	38.6%	8.6%	35.7%	35.2%	45.1%	26.3%
FN removal	481,479	17,554	17,899	13.0%	9.7%	23.5%	8.6%	31.3%	35.2%	32.6%	26.3%
KW Filter	459,353	16,862	17,188	13.4%	10.1%	23.7%	8.8%	28.0%	28.3%	32.1%	24.6%
Uncertainty Filter	429,774	15,402	15,829	12.5%	10.1%	23.7%	4.5%	26.7%	28.3%	32.1%	19.5%
Similarity Filter	401,717	14,339	14,740	12.8%	10.5%	23.7%	4.5%	25.2%	24.1%	32.1%	19.5%

Table 7: Each row describes the state of the dataset upon application of the given filter. The share of wrong labels and bad explanations is only representative of the training split. The first row describes the state of the dataset in its raw form, i.e., before any of the automatic filtering steps. The second row describes the state of the datasets upon application of the false neutral (FN) removal filter, etc.

training jointly can be advantageous, because we backpropagate the explanation loss into the task model M_T, but this also comes at the risk of averse effects on the optimization [14]. The authors of the PJ-X model mentioned that they tried both training approaches, but they do not specify which one worked best. Wu and Mooney [46] only trained separately. It should be noted that PJ-X and FME were both solely run on VQA-X, where a much larger dataset VQA v2 exists for task T. They pre-train M_T separately on this dataset, and it could be argued that, when training jointly, M_T runs the risk of becoming worse by overfitting on the smaller dataset VQA-X. For e-SNLI-VE and VCR, no such pre-training dataset exists. In this work, we train both jointly and separately for every model.

C.3. Reproducing Previous Results

In this work, we reproduced three different models. The code for RVT was publicly available and we only had to add a classifier that is suited for the input type of RVT. The code of PJ-X is also publicly available, albeit in an outdated version of the Caffe framework, and therefore we translated it into Pytorch. For FME no code is available and thus we re-implemented their model (as much as possible) according to the instructions given in the paper [46]. In Table 8 we show that the NLG metrics of our re-implementations come very close to those reported in the original papers.

For PJ-X and FME, we had to make a few minor deviations from the original implementations. To address issues with the gradients (vanishing and destabilizing) in PJ-X, we changed the L2 normalization to layer normalization [7] in the decoder, and added gradient clipping with a threshold of 0.1. FME was re-implemented in contact with the first author of the original paper. We re-implemented their "base" model, which leaves out some of their model extensions. This is motivated by the fact that these extensions either did not lead to performance increases for us (their $\mathcal{L}_F$ loss) or are difficult to reproduce from the descriptions in the paper (their dataset filter $\mathcal{F}$). For the sake of standardization, we use a ResNet-101 as feature extractor for both models. We also tried a ResNet-152, but this had little effect on our results.

C.4. Hyperparameters

In total, we have four models and three datasets. For PJ-X and FME, we choose the same hyperparameters as the authors across all datasets. For PJ-X, we also experimented with larger learning rates, as we experienced convergence issues. For RVT and e-UG, we conducted grid search on three batch sizes, three learning rates, and three ways to combine the loss. We compared dynamic weight loss [32] (with two loss temperatures $T = 2$ and $T = 0.5$) with simply adding both losses. However, this did not affect our results enough to warrant the increase in complexity. We selected the best configuration on VQA-X and then used these settings to train on e-SNLI-VE and VCR. For BERT on VCR, we had to use a higher batch size (128), as the results would not have converged otherwise. The final hyperparameters for all four models are reported in Table 9.

An additional overview of the differences between the models is given in Table 10.

C.5. Adaptations for VCR

To accommodate for the multiple-choice nature of task T, we adapt the architectures accordingly. For UNITER, we follow the original paper and formulate multiple-choice as a binary classification of question-image-answer tuples as True or False. The final answer is determined through a softmax of the four True scores. For PJ-X and FME, we follow the approach in the original VCR paper and obtain the logit for response j via the dot product of the final representation of the model and the final hidden state of the LSTM encoding of the response r^j [50]. For RVT, we use BERTFORMULTIPLECHOICE from the transformers library [45].

Model		BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
PJ-X [37]	*Original*	19.8	18.6	44.0	73.4	15.4
	Ours	20.1	18.3	43.0	71.8	15.3
FME [46]	*Original*	23.5	19.0	46.2	81.2	17.2
	Ours	20.8	19.2	44.8	77.9	16.7

Table 8: A comparison (under the same settings) of automatic NLG metrics on VQA-X between our re-implementations (*Ours*) of PJ-X and FME and the results reported in the papers (*Original*).

	PJ-X	FME	RVT	e-UG
Batch Size	128	128	32* / 64	64
Learning Rate (LR)	7×10^{-4}	5×10^{-4}	5×10^{-5}	2×10^{-5}
Training Type	JOINT*	JOINT*	SEPARATE	JOINT
Loss Combination	$\mathcal{L}_T + \mathcal{L}_E$	$\mathcal{L}_T + \mathcal{L}_E$	N.A.	$\mathcal{L}_T + \mathcal{L}_E$
Optimizer	Adam	Adam	AdamW	AdamW for BERT
LR Scheduler	-	Step decay	Linear w/ warmup	Linear w/ warmup
Tokenization	Word	Word	WordPiece	WordPiece
Max Question Length	23	23	19	19
Max Answer Length	23	40	23	23
Max Explanation Length	40	40	51	51
Decoding	Greedy	Greedy	Greedy	Greedy

Table 9: Hyperarameters used for the different models across all datasets. $\mathcal{L}_T$ and $\mathcal{L}_E$ are the task loss and explanation loss, respectively. For RVT, the task batch size for VCR is 128, as 32 did not lead to convergence. For PJ-X and FME, we trained M_T and M_E separately on VQA-X.

Model M	Vision Backbone	VL Model M_T	Explanation Model M_E	M_E Input
PJ-X	ResNet-101	MCB	LSTM (a)	image features, question, answer
FaiMu	ResNet-101	UpDown	LSTM (b)	image features, question, answer
RVT	Faster R-CNN	BERT	GPT-2	object tags, question, answer
e-UG	Faster R-CNN	UNITER	GPT-2	contextualized embeddings of image-question pair, question, answer

Table 10: Summary of the model differences.

D. Human Evaluation Framework

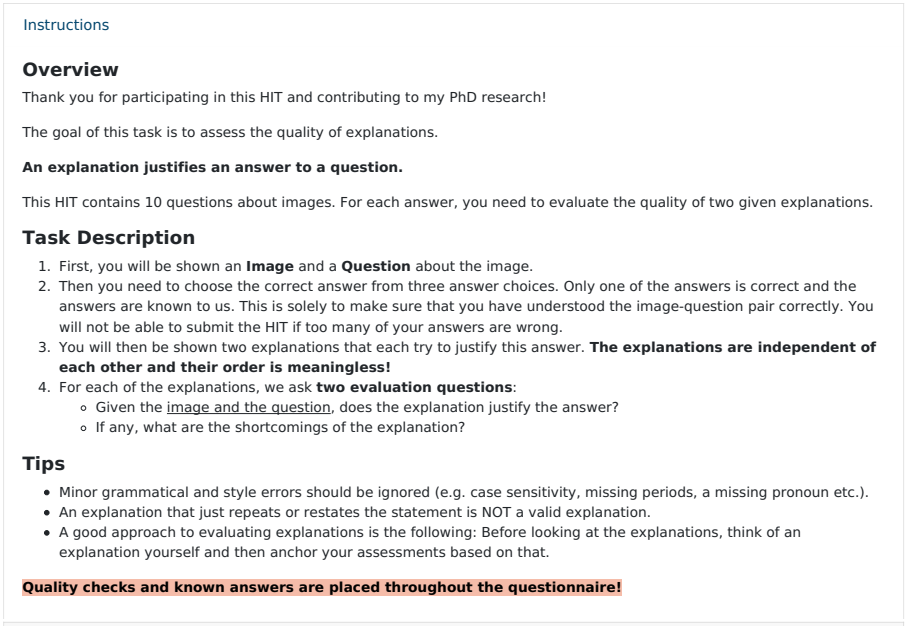
An example of the instructions that were shown to the MTurk annotators can be seen in Figure 14. The interface through which the annotators evaluated the explanations is displayed in Figure 15. The cost to evaluate *one* model on *one* dataset is 108-117$.

E. Results

In this section, we present a benchmark evaluation with automatic NLG metrics (E.1), extended results on e-SNLI-VE performance (E.2) and different ways to compute the e-ViL S_E score (E.4).

E.1. Automatic NLG Metrics

We report the automatic NLG scores in Table 11. Those are computed for all the explanations from the test sets where the predicted answer was correct. A quick observation is that the human evaluation results are not always reflected by the automatic metrics. For example, on the VCR dataset, FME, and not e-UG, obtains the highest S_E score when using automatic NLG metrics. Some tendencies are reflected nonetheless, such as the fact that e-UG is the best model



Instructions

Overview

Thank you for participating in this HIT and contributing to my PhD research!

The goal of this task is to assess the quality of explanations.

An explanation justifies an answer to a question.

This HIT contains 10 questions about images. For each answer, you need to evaluate the quality of two given explanations.

Task Description

1. First, you will be shown an **Image** and a **Question** about the image.
2. Then you need to choose the correct answer from three answer choices. Only one of the answers is correct and the answers are known to us. This is solely to make sure that you have understood the image-question pair correctly. You will not be able to submit the HIT if too many of your answers are wrong.
3. You will then be shown two explanations that each try to justify this answer. **The explanations are independent of each other and their order is meaningless!**
4. For each of the explanations, we ask **two evaluation questions**:
   - Given the image and the question, does the explanation justify the answer?
   - If any, what are the shortcomings of the explanation?

Tips

- Minor grammatical and style errors should be ignored (e.g. case sensitivity, missing periods, a missing pronoun etc.).
- An explanation that just repeats or restates the statement is NOT a valid explanation.
- A good approach to evaluating explanations is the following: Before looking at the explanations, think of an explanation yourself and then anchor your assessments based on that.

Quality checks and known answers are placed throughout the questionnaire!

Figure 14: A snapshot of the instructions that were provided to the annotators that evaluated the explanations.

overall and that e-UG consistently outperforms RVT (albeit by a small margin).

Question-only GPT-2. In order to verify our intuition that the object labels used by RVT provide very little information about the image, we trained GPT-2 that only conditions on the question and answer, ignoring the image (called *GPT-2 only* in Table 11). Without having any image input, this model closely shadows the performance of RVT on most metrics. RVT is still slightly better in most cases, indicating that the object labels do provide some minor improvement. This suggests that RVT is not able to use visual information effectively and learns the explanations mostly based off spurious correlations and not based on the image.

E.2. Detailed Results for e-SNLI-VE

Here, we provide more detailed results on our newly released e-SNLI-VE dataset. We break down the task accuracy and explanation scores by the three different classes (see Table 12). For all models, we observe significantly lower accuracies and explanation scores for the neutral class. There are two potential explanations for this. First, the neutral class can be harder to identify than the other classes. In image-hypothesis pairs, entailment and contradiction examples can sometimes be reduced to more straightforward yes/no classifications of image descriptions. For the neutral class, there always needs to be some reasoning involved to decide whether the image does (not) contain enough evidence to neither indicate entailment nor contradiction. A second reason is that, despite our best efforts to clean the dataset, the neutral class is still more noisy and less represented in the training data.

E.3. Statistical Analysis of the S_E Score

To ensure high quality of our results, we had a number of in-browser checks that prevented the annotators from submitting the questionnaire when their evaluations seemed of poor quality. Checks include making sure that they cannot simultaneously say that an explanation is insufficient (they select the *No* or *Weak No* option described in Section 4) and has no shortcomings, or that it is optimal (they select *Yes* option), but has shortcomings. We also experimented with further post-hoc cleaning measures (such as verifying that they evaluated the ground-truth favorably or did not always choose similar answers), but they had a negligible impact and thus were disregarded.

Our MTurk sample consists of 19,194 evaluations, half of which are for ground-truth explanations, and the other half for model generated explanations. We obtain evaluations for 264 to 299 unique question-image pairs for every model-dataset combination, leaving us with explanations missing for only 3.3% of questions. There are 82.1 evaluations per annotator on average ($SD = 170.1$), ranging from 16 to 1,244 with a median of 34. After pooling annotations of the same explanation, 6,494 annotations remain (887 to 897 for the evaluations generated by each model).

In Figure 16, we add standard errors to the numerical S_E scores given in Table 3. This figure confirms that e-UG uniformly outperforms the other models.

To further investigate the robustness of the e-ViL benchmark, we do a statistical analysis of our S_E scores by using a Linear Mixed Model (LMM) that predicts S_E from the

	e-ViL Scores (auto)			n-gram Scores								Learned Sc
VQA-X	S_O	S_T	S_E	B1	B2	B3	B4	R-L	MET.	CIDEr	SPICE	BERTScore
PJ-X [37]	32.1	76.4	42.1	**57.4**	42.4	30.9	22.7	**46.0**	19.7	82.7	17.1	84.6
FME [46]	33.0	75.5	43.7	59.1	**43.4**	**31.7**	23.1	47.1	20.4	**87.0**	18.4	85.2
RVT [34]	26.8	68.6	39.1	51.9	37.0	25.6	17.4	42.1	19.2	52.5	15.8	85.7
GPT-2 only	N.A.	N.A.	37.8	51.0	36.4	25.3	17.3	41.9	18.6	49.9	14.9	85.3
e-UG	**36.5**	**80.5**	**45.4**	57.3	42.7	31.4	**23.2**	45.7	**22.1**	74.1	**20.1**	**87.0**
VCR												
PJ-X [37]	7.2	39.0	18.4	21.8	11.0	5.9	3.4	20.5	16.4	19.0	4.5	78.4
FME [46]	17.0	48.9	**34.8**	**23.0**	**12.5**	**7.2**	**4.4**	**22.7**	**17.3**	27.7	**24.2**	**79.4**
RVT [34]	15.5	59.0	26.3	18.0	10.2	6.0	3.8	21.9	11.2	30.1	11.7	78.9
GPT-2 only	N.A	N.A	26.3	18.0	10.2	6.0	3.8	22.0	11.2	30.6	11.6	78.9
e-UG	**19.3**	**69.8**	27.6	20.7	11.6	6.9	4.3	22.5	11.8	**32.7**	12.6	79.0
e-SNLI-VE												
PJ-X [37]	26.5	69.2	38.4	29.4	18.0	11.3	7.3	**28.6**	14.7	72.5	24.3	79.1
FME [46]	29.9	73.7	40.6	**30.6**	19.2	12.4	8.2	29.9	15.6	83.6	26.8	79.7
RVT [34]	31.7	72.0	44.0	29.9	19.8	13.6	**9.6**	27.3	18.8	81.7	32.5	81.1
GPT-2 only	N.A.	N.A.	43.6	29.8	19.7	13.5	9.5	27.0	18.7	80.4	32.1	81.1
e-UG	**36.0**	**79.5**	**45.3**	30.1	**19.9**	**13.7**	**9.6**	27.8	**19.6**	**85.9**	**34.5**	**81.7**

Table 11: Automatic NLG metrics for all model-dataset pairs. The S_E based on automatic NLG metrics is the harmonic mean that was used to select the best model during validation. B1 to B4 stand for BLEU-1 to BLEU-4, R-L for ROUGE-L, and MET for METEOR.

	Entailment			Neutral			Contradiction		
	Acc.	MET.	BERTS.	Acc.	MET.	BERTS.	Acc.	MET.	BERTS.
PJ-X	74.4	14.0	79.2	61.5	12.4	77.4	72.8	15.9	79.3
FME	77.3	15.1	79.8	67.3	13.5	77.9	77.2	16.3	79.8
RVT	74.6	17.9	81.3	63.3	**19.0**	80.7	79.4	19.4	81.4
e-UG	**80.3**	**19.6**	**81.6**	**71.7**	18.5	**80.9**	**87.5**	**20.9**	**82.6**

Table 12: Class-wise results on e-SNLI-VE for the different models. NLG metrics are only shown for METEOR and BERTScore, as those correlate most with human judgement.

model-dataset pairs, with model as fixed factor and dataset as random effect. LMM predicts the evaluations with the Likelihood-Ratio-Test of the fixed effect being significant, with $\chi^2(3) = 37.462, p < 0.001$. To gain better insight, we performed post-hoc pairwise contrasts, which indicate that e-UG significantly outperforms the remaining models, with $p < 0.001$. Further, RVT outperforms PJ-X significantly, with $p = 0.007$. The significance level was adjusted for a family-wise type I error rate of $\alpha = 0.05$ using Bonferroni-Holm adjustments.

E.4. Alternative S_E Scores

The nature of our human evaluation questionnaire allows for multiple ways to compute the e-ViL score S_E of the generated explanations. The key differences between the scoring methods are on how to pool the up-to-three evaluations we have for each explanation, and how to compute the overall numerical value. In the main paper, we compute S_E by mapping the four evaluation choices to numerical values, then taking the average for every explanation in the sample and then the sample average to get our S_E score. Below, we propose two alternative ways to compute S_E. While they

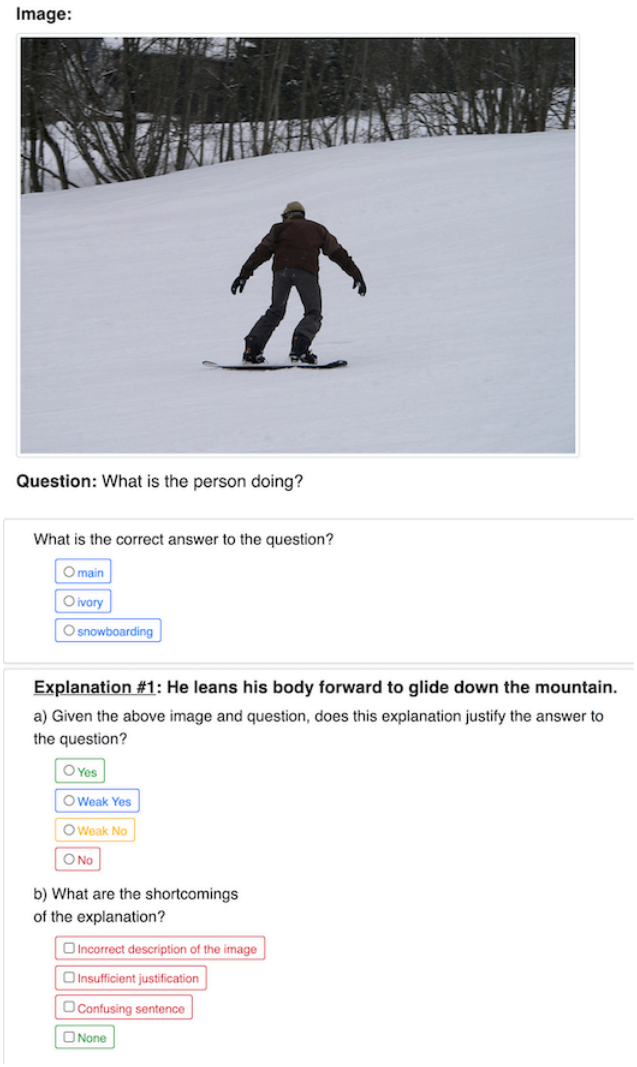


Figure 15: A snapshot of the interface through which annotators evaluated the explanations.

lead to different values, the performance differences between our models remain relatively similar.

E.4.1 Median Pooling

In median pooling, we obtain the score for each explanation by taking the median of its up-to-three ordinal evaluations (as opposed to taking a numerical average). We always interpolate with rounding off, meaning that the median of (*Yes*, *Weak Yes*) $\mapsto$ *Weak Yes* and (*Yes*, *No*) $\mapsto$ *Weak No*. This allows us to plot the distribution of *No*, *Weak No*, *Weak Yes*, and *Yes* for every model-dataset pair, as displayed in Figure 18.

We observe that e-UG performs better across all datasets, with RVT following in second place for the VCR and VQA-X datasets. The differences between the PJ-X, FME and RVT are relatively small.

We analyse our results using a Cumulative Link Mixed Model (CLMM) with a logit link and flexible thresholding. We predict annotator responses using the dataset as random effect and the VL-NLE model as fixed effect. We find that the model significantly influences ratings, as suggested by the Likelihood-Ratio-Test, $\chi^2(2) = 42.4, p < 0.001$, when comparing the full model to a nested statistical model that is merely based on the dataset as predictor. The model predictor is dummy-coded with e-UG as reference class , which enables us to interpret the model's coefficients in the statistical test as pairwise contrasts of all other models towards e-UG. All coefficients have p-values $p < 0.001$, indicating the e-UG significantly outperforms all other models.

E.4.2 Comparative S_E Score

We also designed a comparative score, for which we do not map our questionnaire evaluation options (*No*, *Weak No*, *Weak Yes*, and *Yes*) to numerical values, but instead compare them to the evaluation of the ground-truth. For every image-question pair, the annotator has to evaluate both the ground-truth and the generated explanation, without knowing which is which. This enables us to see, for every generated explanation, if it was deemed equally good, better, or worse than the ground-truth. This mimics the approach in Park et al. [37] and Wu and Mooney [46], where annotators were explicitly asked if the generated explanation was worse, equally good, or better than the ground-truth. An advantage of this method is that we can seamlessly incorporate the criticalness of each annotator. The disadvantage is that we do not get *absolute* measurements of the quality of the explanations.

The generated explanation gets the score 1 if it is as good or better than the ground-truth, and otherwise 0. We pool the comparative score via median pooling with rounding off.

Figure 17 displays the comparative score. We can observe that e-UG scores are strongest across all datasets, while the other three models are performing similarly, except on the VCR dataset, where PJ-X performs worse than the other models.

For our statistical analysis, we fit a generalized linear mixed model (GLMM) on the full unpooled annotation set predicting the whether an explanation was rated positively (compared to the ground-truth) using the dataset and annotator as random effects and the VL-NLE model as fixed effect. We utilise a logit link. The model parameter significantly predicts the evaluations, with $\chi^2(3) = 67.366, p < 0.001$. Post-hoc tests (Tukey contrasts with Bonferroni-Holm adjusted significance) show that the e-UG outperforms all other models, with $p < 0.001$, and that RVT outperforming PJ-X at $p = 0.011$. All other pairwise comparisons were not significant. Extending the model to include ground-truth

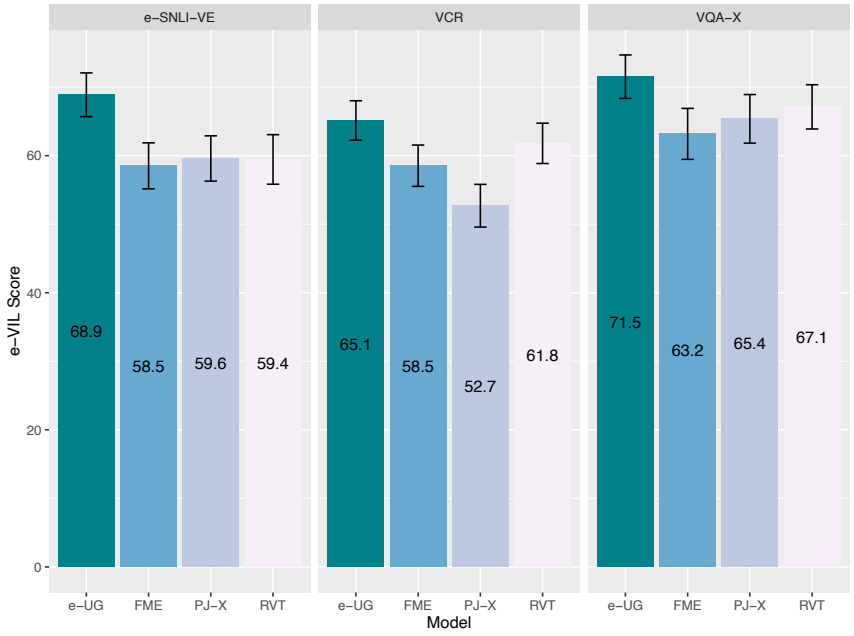


Figure 16: Human evaluation framework: e-ViL scores S_E. This plot shows the main e-ViL scores (based on numerical average) for the different model-dataset pairs. Error bars show $\pm 2\text{SD}/\sqrt{n}$ for each group.

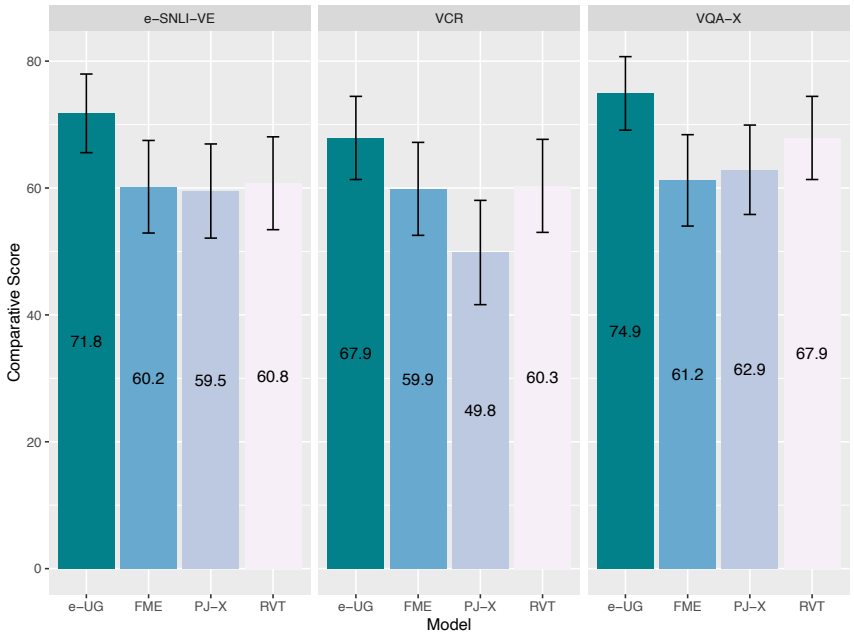


Figure 17: Human evaluation framework: Comparative scores. This figure displays the comparative scores (with respect to the ground-truth) of the explanations for the different model-dataset pairs. Error bars show $\pm 2\text{SD}/\sqrt{n}$ for each group.

explanations as a model category also demonstrates that all model-generated explanations were evaluated significantly worse than the ground-truth explanations. We conclude that the e-UG outperforms all other models, whereas performance differences between them are rather small, replicating our findings from the alternative analyses.

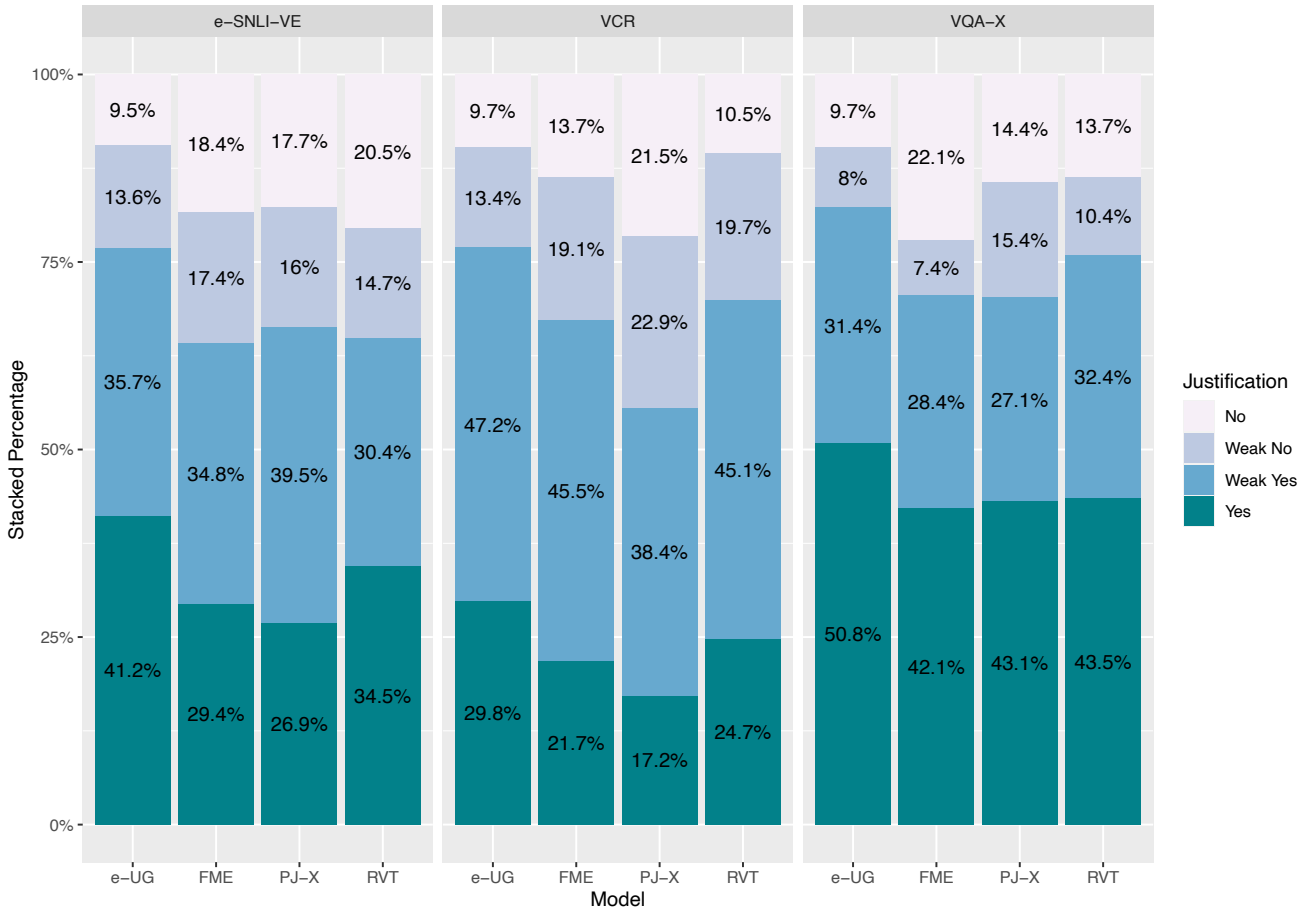


Figure 18: Human evaluation framework: Ordinal representation of the evaluations. Median responses for each question-image pair given by participants to the evaluation question question "Given the image and the question/hypothesis, does the explanation justify the answer?".