

# e-ViL: A Dataset and Benchmark for Natural Language Explanations in Vision-Language Tasks

Maxime Kayser<sup>1\*</sup> Oana-Maria Camburu<sup>1</sup> Leonard Salewski<sup>2</sup> Cornelius Emde<sup>1</sup>  
Virginie Do<sup>1\*\*</sup> Zeynep Akata<sup>2,3,4</sup> Thomas Lukasiewicz<sup>1</sup>

<sup>1</sup>Department of Computer Science, University of Oxford    <sup>2</sup>University of Tübingen

<sup>3</sup>Max Planck Institute for Intelligent Systems    <sup>4</sup>Max Planck Institute for Informatics

## Abstract

*Recently, there has been an increasing number of efforts to introduce models capable of generating natural language explanations (NLEs) for their predictions on vision-language (VL) tasks. Such models are appealing, because they can provide human-friendly and comprehensive explanations. However, there is a lack of comparison between existing methods, which is due to a lack of re-usable evaluation frameworks and a scarcity of datasets. In this work, we introduce e-ViL and e-SNLI-VE. e-ViL is a benchmark for explainable vision-language tasks that establishes a unified evaluation framework and provides the first comprehensive comparison of existing approaches that generate NLEs for VL tasks. It spans four models and three datasets and both automatic metrics and human evaluation are used to assess model-generated explanations. e-SNLI-VE is currently the largest existing VL dataset with NLEs (over 430k instances). We also propose a new model that combines UNITER [15], which learns joint embeddings of images and text, and GPT-2 [38], a pre-trained language model that is well-suited for text generation. It surpasses the previous state of the art by a large margin across all datasets. Code and data are available here: <https://github.com/maximek3/e-ViL>.*

## 1. Introduction

Deep learning models achieve promising performance across a variety of tasks but are typically black box in nature. There are several arguments for making these models more explainable. For example, explanations are crucial in establishing trust and accountability, which is especially relevant in safety-critical applications such as healthcare or autonomous vehicles. They can also enable us to better understand and correct the learned biases of models [5].

Explainability efforts in vision tasks largely focus on highlighting relevant regions in the image, which can be achieved via tools such as saliency maps [1] or attention maps [47]. Our work focuses on natural language explanations (NLEs), which aim to explain the decision-making process of a model via generated sentences. Besides being easy to understand for lay users, NLEs can explain more complex and fine-grained reasoning, which goes beyond highlighting the important image regions. We compare different models that generate NLEs for vision-language (VL) tasks, i.e., tasks where the input consists of visual and textual information, such as visual question-answering (VQA).

NLEs for VL tasks (VL-NLE) is an emerging field, and only few datasets exist. Moreover, existing datasets tend to be relatively small and unchallenging (e.g., VQA-X [37]) or noisy (e.g., VQA-E [29]). Another limitation of the VL-NLE field is that there is currently no unified evaluation framework, i.e., there is no consensus on how to evaluate NLEs. NLEs are difficult to evaluate, as correct explanations can differ both in syntactic form and in semantic meaning. For example, “Because she has a big smile on her face” and “Because her team just scored a goal” can both be correct explanations for the answer “Yes” to the question “Is the girl happy?”, but existing automatic natural language generation (NLG) metrics are poor at capturing this. As such, the gold standard for assessing NLEs is human evaluation. Past work have all used different approaches for human evaluations, and therefore no objective comparison exists.

In this work, we propose five main contributions to address the lack of comparison between existing work. (1) We propose e-ViL, the first benchmark for VL-NLE tasks. e-ViL spans across three datasets of human-written NLEs, and provides a unified evaluation framework that is designed to be re-usable for future works. (2) Using e-ViL, we compare four VL-NLE models. (3) We introduce e-SNLI-VE, a dataset of over 430k instances, the currently largest dataset for VL-NLE. (4) We introduce a novel model, called e-UG,

\*Corresponding Author: maxime.kayser@cs.ox.ac.uk

\*\*Now at Université Paris-Dauphine, PSL, and Facebook AI Research.

which surpasses the state of the art by a large (and significant) margin across all three datasets. (5) We provide the currently largest study on the correlation between automatic NLG metrics and human evaluation of NLEs.

## 2. Related Work

**Explainability in Computer Vision.** Common approaches to explainability of deep learning methods in computer vision are saliency maps [1], attention maps [47], and activation vectors [25]. Saliency and attention maps indicate where a model looks. This may tell us what regions of the image are most important in the decision-making process of a model. Activation vectors are a way to make sense of the inner representation of a model, e.g., by mapping it to human-known concepts. However, these approaches often cover only a fraction of the reasoning of a model. On the contrary, NLEs can convey higher-order reasoning and describe complex concepts. For example, in Figure 1, highlighted image regions or weights for different concepts would not be sufficient to explain the answer. Additionally, it has been shown that numerical or visual explanatory methods, in some cases, can be confusing even for data scientists [24] and can pose problems even for explaining trivial models [13].

**NLEs.** First adoptions of NLEs have been in image classification [23] and were further extended to self-driving cars [26], VQA [37], and natural language processing [11; 39; 9; 6; 27; 35; 28; 12]. The most important works in VL-NLE [37; 46; 34] are included in this benchmark.

**VL-NLE Datasets.** Existing models learn to generate NLEs in a supervised manner and, therefore, require training sets of human-written explanations. Besides the image classification datasets ACT-X [37] and CUB [43; 23], and the video dataset BDD-X [26], there are currently three VL datasets with NLEs. The VQA-X dataset [37] was introduced first and provides NLEs for a small subset of questions from VQA v2 [4]. It consists of 33k QA pairs (28k images). However, many of the NLEs in VQA-X are trivial and could be guessed without looking at the image. For example, “because she is riding a wave on a surfboard” is an NLE for the answer “surfing” to the question “What is the woman in the image doing?” that can easily be guessed from the answer, without looking at the image (more examples are given in Figure 6). VQA-E [29] is another dataset that also builds on top of VQA v2. However, its explanations are gathered in an automatic way and were found to be of low quality by Marasović et al. [34], where the model-generated explanations obtain a human evaluation accuracy<sup>1</sup> that is only 3% short of the VQA-E ground-truth explanations (66.5%), suggesting that the dataset is essentially solved. It is therefore not used in our benchmark. Finally, the VCR dataset [50]

<sup>1</sup>Percentage of explanations that, given the image and question, support the predicted answer.

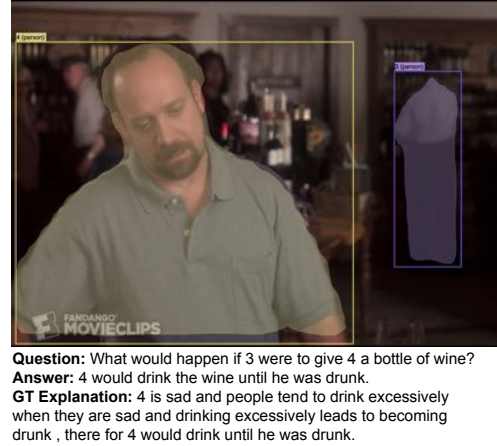


Figure 1: VCR images require commonsense reasoning that often goes beyond the visual content of the image.

provides NLEs for VQA instances that require substantial commonsense knowledge (see Figure 1). The questions are challenging, and therefore both the answers and NLEs are given in the form of multiple-choice options.

Our proposed dataset, e-SNLI-VE, extends the range of VL-NLE datasets and addresses some of the prior limitations. It contains over 430k instances for which the explanations rely on the image content (see examples in Figure 6). We will describe the dataset in more detail in Section 3.

**Evaluations and Comparisons.** Evaluating NLG is challenging and a much-studied field [19]. Evaluating NLEs is even more difficult, as sentences may not only differ in their syntactic form but also in their semantic meaning, e.g., there can be several different reasons why a sentence contradicts an image. For this reason, current automatic NLG metrics, such as the BLEU score [36], do not perform well in evaluating NLEs [11]. Hence, several works have used human evaluation to assess their generated explanations [37; 46; 34; 17]. However, they all used different evaluation rules, preventing one from being able to compare existing VL-NLE models. The main differences lie in the datasets used, the questions asked to annotators, whether the assessment is absolute or based on a ranking, and the formula used to calculate the final score. We select the best practices from existing evaluation schemes and develop a unified and re-usable human evaluation framework for VL-NLE.

## 3. The e-SNLI-VE Dataset

We introduce e-SNLI-VE, a large-scale dataset for visual-textual entailment with NLEs. We built it by merging the explanations from e-SNLI [11] and the image-sentence pairs from SNLI-VE [48]. We use several filters and manual relabeling steps to address the challenges that arise from



**Hypothesis:** A man and woman inside a church.  
**Textual premise:** A man and woman getting married.  
**Original label:** Neutral  
**Caption #2:** A man and woman that is holding flowers smile in the sunlight.  
**Caption #4:** A happy couple enjoying their open air wedding.

Figure 2: The original label of the textual premise-hypothesis pair in SNLI is neutral. However, by considering alternative captions describing the same image (#2 and #4), we can deduct that the neutral label is false.

merging these datasets. The validation and test sets were relabelled by hand. The dataset is publicly available<sup>2</sup>.

### 3.1. Correcting SNLI-VE

In SNLI-VE [48], an image and a textual hypothesis are given, and the task is to classify the relation between the image-premise and the textual hypothesis. The possible labels are *entailment* (if the hypothesis is true, given the image), *contradiction* (the hypothesis is false, given the image), or *neutral* (if there is not enough evidence to conclude whether the hypothesis is true or false). SNLI-VE builds off the SNLI [10] dataset, by replacing textual premises with Flickr30k images [49]. This is possible, because the textual premises in SNLI are caption sentences of those images. However, this replacement led to labeling errors, as an image typically contains more information than a single caption describing it. Especially for the neutral class, a caption may not have enough evidence to suggest entailment or contradiction, but the corresponding image does (see Figure 2). On a manually evaluated subset of 535 samples, we found a 38.6% error rate among the neutral labels. This subset will be used below to evaluate the effectiveness of our filters. Error rates for entailment and contradiction are reported to be under 1% [48], hence we focus only on correcting the neutral instances.

In the validation and test sets, we relabeled the neutral examples using Amazon Mechanical Turk (MTurk). To ensure high-quality annotations, we used a series of quality control measures, such as in-browser checks, inserting trusted examples, and collecting three annotations per instance. In total, 39% of the neutral labels were changed to entailment or contradiction. The label distribution shifted from uniform

to Ent/Neut/Cont of 39%/20%/41% and 39%/21%/40% for the validation and test sets, respectively.

For the training set, we propose an automatic way to remove false neutrals. We discovered that the five captions that come with each image often provide clues whether the label is indeed neutral, or not. For every image-hypothesis pair  $i$ , we ran a natural language inference model  $m_{\text{nli}}$  on each caption-hypothesis pair  $p_{i,c}$ , where  $c$  is one of the captions. If the original label of image-hypothesis pair  $i$  is neutral, but  $\sum_c m_{\text{nli}}(p_{i,c})$  indicates with high confidence that the label is not neutral, we deem the label incorrect and removed the instance from the dataset. An example is shown in Figure 2. For  $m_{\text{nli}}$ , we used Roberta-large [33] trained on the MNLI dataset [44]. Instances were removed if  $\sum_c m_{\text{nli}}(p_{i,c})$  exceeded 2.0 for entailment and contradiction classes. On our 535-samples subset, this filter decreased the error of neutral labels from 39% to 24%. When validated against the relabeling on the validation set, the error decreased from 39% to 30%.

### 3.2. Adding Explanations to SNLI-VE

To create e-SNLI-VE, we source explanations from e-SNLI [11], which extends SNLI with human-written NLEs. However, the explanations in e-SNLI are tailored to the textual premise-hypothesis pairs and are therefore not always well-suited for the image-hypothesis pair. After simply merging both datasets, we found that initially 36%, 22%, and 42% of explanations were of low (incorrect), medium (correct, but there is an obvious better choice), and high quality (correct and relevant), respectively. We propose several steps to detect and remove explanations of low and medium quality. The filters were designed to ensure an optimal trade-off between precision and recall (for flagging bad explanations) and with the constraint that the final dataset remains reasonably balanced.

**Re-annotation.** First, we replace the explanations for the neutral pairs in the validation and test sets with new ones, collected via MTurk at the same time as we collected new labels for these subsets. In order to submit the annotation of an image-sentence pair, workers must choose a label, highlight words in the hypothesis, and use at least half of the highlighted words in the explanation.

**Keyword Filter.** Next, we use keyword filtering to detect explanations that make reference to a linguistic feature of the textual premise. The keywords, which we manually defined, are “synonym”, “mention”, “rephrasing”, “sentence”, “way to say”, and “another word for”. The keyword filter removed 4.6% of all instances, and our 535-samples subset suggests that *all* filtered explanations were indeed of low quality.

**Similarity Filter.** We noticed that the share of low-quality explanations is highest for entailment examples. This happens frequently when the textual premise and hypothesis are almost identical, as then the explanation often just repeats

<sup>2</sup><https://github.com/maximek3/e-ViL>

	Train	Validation	Test
# Image-Hypothesis pairs (# Images)	401,717 (29,783)	14,339 (1,000)	14,740 (1,000)
Label distribution (C/N/E, %)	36.0 / 31.3 / 32.6	39.4 / 24.0 / 36.6	38.8 / 25.8 / 35.4
Mean hypothesis length (median)	7.4 (7)	7.3 (7)	7.4 (7)
Mean explanation length (median)	12.4 (11)	13.3 (12)	13.3 (12)

Table 1: e-SNLI-VE summary statistics. C, N, and E stand for Contradiction, Neutral, and Entailment, respectively.

both statements. To overcome this, we removed all examples where the ROUGE-1 score (a measure for sentence similarity [31]) between the textual premise and hypothesis was above 0.57. This reduced the share of low-quality explanations for entailment by 4.2%.

**Uncertainty Filter.** Lastly, we found that image-hypothesis pairs with high uncertainty are correlated with low-quality explanations for contradictions. We define uncertainty as the diversion of the scores from  $m_{\text{nli}}(p_{i,c})$  for the five image captions.  $m_{\text{nli}}$  is the same Roberta-large model that was described above. This filter reduced the share of low-quality explanations for contradiction examples by 5.1%.

The final e-SNLI-VE dataset statistics are displayed in Table 1. An additional evaluation of e-SNLI-VE by external annotators, and a comparison with existing VL-NLE datasets, is provided in Table 2. The results indicate that the quality of the e-SNLI-VE ground-truth explanations is not far off the human-annotated VQA-X and VCR datasets. Qualitative examples and a more detailed rundown of our filtering methods are in Appendix B.

## 4. The e-ViL Benchmark

In this section, we introduce the VL-NLE task, describe how explanations are evaluated in e-ViL, and describe the datasets covered in our benchmark.

### 4.1. Task Formulation

We denote a module that solves a VL task as  $M_T$ , which takes as input visual information  $V$  and textual information  $L$ . Its objective is to complete a task  $T$  where the outcome is  $a$ , i.e.,  $M_T(V, L) = a$ . An example of a VL task is VQA, where  $V$  is an image,  $L$  is a question, and  $T$  is the task of providing the answer  $a$  to that question. We extend this by an additional task  $E$ , which requires an NLE  $e$  justifying how  $V$  and  $L$  lead to  $a$ , solved by the module  $M_E(V, L) = e$ . The final model  $M$  then consists of  $M_T$  and  $M_E$ . Thus,  $M = (M_T, M_E)$  and  $M(V, L) = a, e$ .

### 4.2. Datasets

Our benchmark uses the following three datasets, which vary in size and domain. Examples are shown in Figure 6 in the appendix.

**e-SNLI-VE.** Our proposed e-SNLI-VE dataset has been described in Section 3.

**VQA-X.** VQA-X [37] contains human written explanations for a subset of questions from the VQA v2 dataset [21]. The image-question pairs are split into train, dev, and test with 29.5k, 1.5k, and 2k instances, respectively. The task  $T$  is formulated as a multi-label classification task of 3,129 different classes. One question can have multiple possible answers.

**VCR.** Visual Commonsense Reasoning (VCR) is a VL dataset that asks multiple-choice (single answer) questions about images from movies [50]. In addition to four answer options, it also provides four NLEs options, out of which one is correct. For the purpose of our proposed VL-NLE task, we reformulate it as an explanation generation task. As the test set for VCR is not publicly available, we split the original train set into a train and dev set, and use the original validation set as test set. The splits are of size 191.6k, 21.3k, and 26.5k, respectively.

**Human Evaluation of Datasets.** In our benchmark experiments (Section 5), human annotators evaluate the ground-truth explanations of all three datasets. For each dataset, 300 examples are evaluated by 12 annotators each, resulting in 3,600 evaluations. The results in Table 2 show that e-SNLI-VE comes close to the manually annotated datasets VCR and VQA-X (82.8% explanations with *yes* or *weak yes* vs. 87.9% and 91.4%). Besides the use of effective, but imperfect, automatic filters, another explanation for the higher share of noise is the trickiness (out of a 100 human re-annotated explanations for neutral examples, we found that 17% had flaws, identical to the share of (weak) no in Table 2) and ambiguity (when three of us chose the labels for a set of 100 image-hypothesis pairs, we only had full agreement on 54% of examples) inherent in the e-SNLI-VE task.

	No	Weak No	Weak Yes	Yes
e-SNLI-VE	10.3%	6.9%	27.7%	55.1%
VQA-X	4.1%	4.5%	25.1%	66.3%
VCR	6.9%	5.2%	36.6%	51.3%

Table 2: Human evaluation of the ground-truth explanations for the three datasets used in e-ViL. The question asked was: “Given the image and the question/hypothesis, does the explanation justify the answer?”.

### 4.3. Evaluation

**Evaluation Scores.** We define separate evaluation scores  $S_T$ ,  $S_E$ , and  $S_O$  for  $M_T$ ,  $M_E$ , and  $M$ , respectively.  $S_T$  is the metric that is defined by the original VL task  $T$ , e.g., label accuracy for e-SNLI-VE and VCR, and VQA accuracy for VQA-X. We define  $S_E$  as the average explanation score of the examples for which the answer  $a$  was predicted correctly. In line with previous work [37; 46; 34], we for now assume the simplified scenario that an explanation is always false when it justifies an incorrect answer. The explanation score can be any custom human or automatic metric. Due to the limitations of current automated NLG metrics for evaluating NLEs, we developed a human evaluation framework for computing  $S_E$ , outlined in the paragraph below. Finally, we want  $S_O$  to summarize the performance of a model on both tasks  $T$  and  $E$ , to give us the overall performance of a VL-NLE model  $M$ . We define  $S_O = S_T S_E$ , which equates to the average of the scores of all explanations, but where we set the score of an explanation to 0 if its associated answer was predicted incorrectly. This can also be viewed the explanation score  $S_E$  multiplied by a coefficient for task performance (accuracy, in most cases). We introduced this measure to avoid giving an advantage to models that purely optimize for generating a few good explanations but neglect the task itself.

**Human Evaluation Framework.** We collect human annotations on MTurk, where we ask the annotators to proceed in two steps. First, they have to solve the task  $T$ , i.e., provide the answer  $a$  to the question. This ensures the annotators think about the question first and enables us to do in-browser quality checks (since we know the answers). We disregard their annotation if they answered the VL task  $T$  incorrectly.

For each explanation, we ask them a simple evaluation question: “Given the image and the question/hypothesis, does the explanation justify the answer?”. We follow Marasović et al. [34] in giving four response choices: *yes*, *weak yes*, *weak no*, and *no*. We map *yes*, *weak yes*, *weak no*, and *no* to the numeric scores of 1,  $\frac{2}{3}$ ,  $\frac{1}{3}$ , and 0, respectively.

We also ask annotators to select the main shortcomings (if any) of the explanations. We observe three main limitations of explanations. First, they can *insufficiently justify the answer*. For example, the sentence “because it’s cloudy” does not sufficiently justify the answer “the sea is not calm”. Second, an explanation can *incorrectly describe the image*, e.g., if a model learned generic explanations that are not anchored in the image. “There is a person riding a surfboard on a wave” is generally a good explanation for the answer “surfing” when asked “what activity is this?”, but the image may actually display a dog surfing. Lastly, the sentences can be *nonsensical*, such as “a man cannot be a man”.

For each model-dataset pair, we select a random sample of 300 datapoints where the model answered the question correctly. Every sample contains only unique images. For

VCR, all movies are represented in the samples. Note that it is not possible to evaluate all models on exactly the same instances, as they do not all answer the same questions correctly. Taking a subset of examples where *all* models answered correctly is disadvantageous for two reasons. First, this makes the benchmark less re-usable, as future methods might not answer the same questions correctly. Second, this would bias the dataset towards the questions that the weakest model answered correctly. However, in order to still maximize the overlap between the samples, we shuffled all the instances in the test sets randomly and then for each model we took the 300 first on which the answer was correct.

We propose several measures to further ensure robustness and re-usability of the framework. In order to account for annotator subjectivity, we evaluate every instance by three different annotators. The final score per explanation is given by the average of all evaluations. In addition, we evaluate one model at a time to avoid potential anchoring effects between models (e.g., the annotator evaluates one model more favorably, because they are influenced by poor explanations from a different model). To implicitly induce a uniform anchoring effect, the annotators evaluate both the ground-truth explanation (which is invariant to the model) and the explanation generated by a model for every image-question pair. They do not know which is which and are not asked to compare them. This implicitly ensures that all evaluations have the same anchor (the ground-truth) and it allows us to compute  $S_E$  in different ways, as outlined in Appendix E.4. Overall, over 200 individual annotators were employed for the benchmark and all of them had to have a 98% prior acceptance rate on MTurk. Finally, we bolster our results with statistical tests in Appendix E.3.

More details and screenshots of our MTurk evaluation can be found in Appendix E. For re-usability, we publicly release the questionnaires used in our benchmark<sup>3</sup>.

## 5. Experimental Evaluation

### 5.1. Models

Existing VL-NLE models follow a common high-level structure (Figure 3). First, a VL model learns a joint representation of the image and language inputs and predicts the answer. The models in this work then condition their explanation on different combinations of the question, image, their joint representation, and the answer. Details on PJ-X [37], FME [46], and RVT [34] are given in Appendix C, as well as in their respective papers.

**e-UG.** Marasović et al. [34] generate convincing explanations, but out of various  $M_T$  modules tested, including complex visual reasoning models, it obtains the best explanation accuracy when using object labels as the sole image information. We address this limitation by proposing e-UG, a model

<sup>3</sup><https://github.com/maximek3/e-ViL>

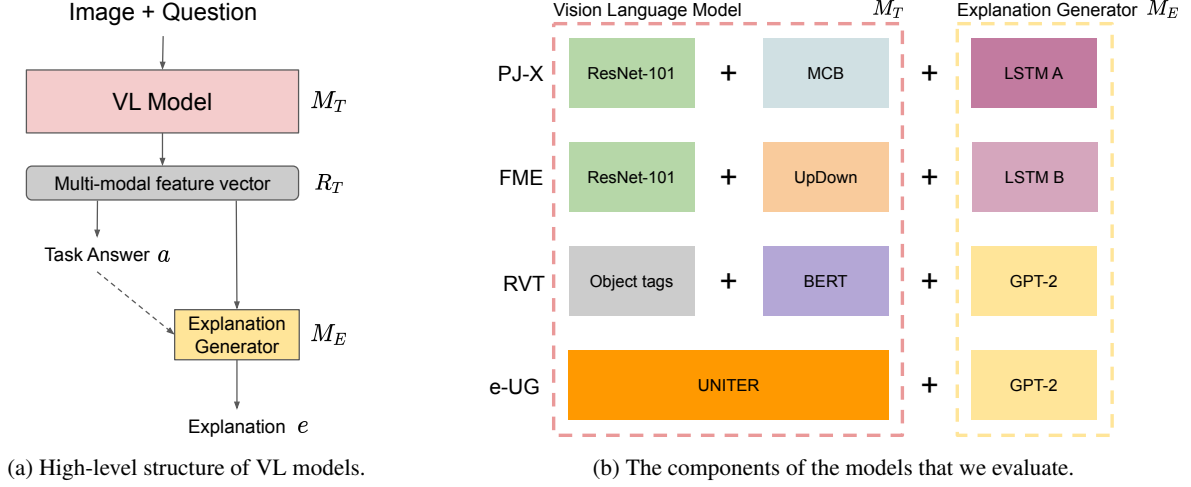


Figure 3: High-level architectures of the models that are included in our benchmark.

that enables a stronger image conditioning by combining GPT-2 with UNITER [15], a powerful transformer-based VL model. The outputs of UNITER are contextualized embeddings of the word tokens and image regions in the image-text pair. Words are embedded by tokenizing them into WordPieces and adding their position embedding. Images are embedded by extracting visual features of regions with Faster R-CNN [40] and encoding their location features. UNITER achieves SOTA on many downstream tasks when fine-tuned on them. For e-UG, we leverage these contextualized embeddings to condition GPT-2 on an efficient representation of the image and question. The embeddings of the image regions and question words are simply prepended to the textual question and predicted answer, and then fed to GPT-2. GPT-2 is a decoder-only architecture that is pre-trained on conventional language modeling and therefore well-suited for language generation [38]. We follow Marasović et al. [34] and do greedy decoding during inference.

## 5.2. Training

All models are trained separately on each dataset. To ensure comparability, image features for PJ-X and FME are obtained from the same ResNet-101 [22] pre-trained on ImageNet, which yields a 2048d feature representation for an image. To account for the small size of VQA-X, the VQA  $M_T$  models were pre-trained on VQA v2 for VQA-X, and trained from scratch for the other two datasets. For UNITER, we follow the pre-training procedures used in the original paper [15]. The object tags in RVT are obtained from a Faster R-CNN that was trained on ImageNet and COCO. For GPT-2, we load the pre-trained weights of the original GPT-2 with 117M parameters [38]. For all models in this work, we experimented with training the  $M_T$  and  $M_E$  modules jointly and separately. More details are given in Appendix C.2.

**Hyperparameters.** Choosing hyperparameters via human evaluation is prohibitively expensive. Instead, we defined a set of automatic NLG metrics that we used to approximate the selection of the best hyperparameters. We define the score of an explanation as the harmonic mean of the BERTScore F1 [51] and NGRAMScore, where we set NGRAMScore as the harmonic mean of the  $n$ -gram NLG metrics ROUGE-L [30], SPICE [2], CIDEr [42], and METEOR [8]. We pick the harmonic mean, as it puts more emphasis on the weaker scores. Further details on the hyperparameters are given in Appendix C.4.

## 5.3. Results

In this section, we highlight the human evaluation results, their correlation with automatic NLG metrics, and the effect that training with explanations has on the performance on task  $T$ . Model performance for automatic NLG metrics, detailed results on e-SNLI-VE, alternative computations of the human evaluation score, and a statistical analysis of the results is provided in Appendix E.

### 5.3.1 Human Evaluation

The explanation scores  $S_E$  obtained from the e-ViL human evaluation framework are displayed in Table 3. Our model e-UG outperforms existing methods on all datasets, with an average  $S_E$  score 5.7 points higher than the second-best model, RVT. Despite leveraging little image information, RVT achieves higher scores than PJ-X and FME on average, reflecting the ability of GPT-2 to learn to generate convincing explanations, without much anchoring on the image. There is still a significant gap between  $S_E$  scores of generated explanations and ground-truth (GT) explanations. For VQA-X,  $S_E$  scores are higher overall, indicating that the dataset is easier. In terms of the overall score  $S_O$ , the gap between e-UG and the rest increases further, as UNITER achieves a



	Overall	VQA-X				e-SNLI-VE			VCR		
	$S_E$	$S_O$	$S_T$	$S_E$	$S_O$	$S_T$	$S_E$	$S_O$	$S_T$	$S_E$	
PJ-X	59.2	49.9	76.4	65.4	41.2	69.2	59.6	20.6	39.0	52.7	
FME	60.1	47.7	75.5	63.2	43.1	73.7	58.5	28.6	48.9	58.5	
RVT	62.8	46.0	68.6	67.1	42.8	72.0	59.4	36.4	59.0	61.8	
e-UG	<b>68.5</b>	<b>57.6</b>	<b>80.5</b>	<b>71.5</b>	<b>54.8</b>	<b>79.5</b>	<b>68.9</b>	<b>45.5</b>	<b>69.8</b>	<b>65.1</b>	
GT	79.3	–	–	84.5	–	–	76.2	–	–	77.3	

Table 3: e-ViL benchmark scores.  $S_O$ ,  $S_T$ , and  $S_E$  are defined in Section 4.3. GT denotes the ground-truth explanations in each dataset. The best results are in bold.

higher performance on VL tasks than the  $M_T$  modules of the other models. In Figure 4, we show an example with the explanations generated by each model. In this example, e-UG is the only model that accurately describes the image and justifies the answer. Additional examples are given in Figure 5 in the appendix.

As a second question, we ask the annotators to select the shortcomings (if any) for every explanation. Results for this are given in Table 5. The most frequent shortcoming is an insufficient justification of the answer. Least frequent, with around 10% prevalence, explanations can be nonsensical (e.g., “a woman is a woman”). All models struggle similarly much with producing explanations that sufficiently justify the answer. e-UG and PJ-X are better at producing coherent sentences. e-UG is significantly superior in terms of the explanations accurately describing the image content. This empirically confirms the effectiveness of our enhanced conditioning on the image. On a dataset level, we see that it is easiest for all models to provide explanations that make grammatical sense and justify the answer on VQA-X, reinforcing the notion that the explanations of VQA-X are easier and less elaborate.

A statistical analysis of our findings are given in Appendix E.

### 5.3.2 Correlation of NLG Metrics with Human Evaluation

To better understand to what extent automatic NLG metrics are able to mirror human judgment of explanations, we compute the Spearman correlation of different NLG metrics with the human evaluation scores. The NLG metrics for the different models are given in Appendix E.1. The human evaluation score is averaged and normalised (across all annotators) for each explanation. We have human evaluation scores for a total of 3,566<sup>4</sup> generated explanations, which makes it the currently largest study on the correlation of NLG metrics with human evaluation in NLEs.

<sup>4</sup>We have 4 models, 3 datasets of 300 examples, therefore 3,600 explanations. However, for 34 of them, all the three annotators answered the question incorrectly.



**Hypothesis:** The people are flying kites at the beach.  
**Answer:** Contradiction  
**RVT:** People can't be riding kites while they are flying kites.  
**PJ-X:** People cannot be flying and flying at the same time.  
**FME:** People cannot be walking and flying kites at the same time  
**e-UG:** People cannot be flying kites while they are standing on a street.  
**GT Explanation:** construction site is different from the beach

Figure 4: Generated explanations for each model on an image-hypothesis pair in e-SNLI-VE.

The results in Table 6 show that BERTScore and ME-TOR exhibit significantly higher correlation with human annotators across all datasets, reaching a maximal value of 0.293, which is a relatively low correlation. The reliability of automatic metrics also differs by dataset. They are highest on VQA-X and lowest on VCR. This could be explained by the fact that explanations in VCR are generally semantically more complex or more speculative (and, therefore, there are more different ways to explain the same thing) than those in VQA-X. It is noteworthy that some  $n$ -gram metrics, such as BLEU, ROUGE, or CIDEr, have no statistically significant correlation with human judgment on VCR.

### 5.3.3 Explanations as Learning Instructions

Training a model jointly on the tasks  $T$  and  $E$  can be viewed as a form of multi-task learning [14]. The explanations  $e$  augment the datapoints of task  $T$  by explaining why an answer  $a$  was given. The module  $M_T$  (which solves task  $T$ ) may bene-

Model	$M_T$ model	VQA-X		SNLI-VE		VCR	
		$M_T$ only	Joint	$M_T$ only	Joint	$M_T$ only	Joint
PJ-X	MCB [18]	N.A.	N.A.	<u>69.7</u>	69.2	38.5	<u>39.0</u>
FME	UpDown [3]	N.A.	N.A.	71.4	<u>73.7</u>	35.7	<u>48.9</u>
e-UG	UNITER [15]	80.0	<u>80.5</u>	79.4	79.5	69.3	<u>69.8</u>

Table 4: Comparison of task scores  $S_T$  (e.g., accuracies) when the models are trained only on task  $T$  vs. when trained jointly on tasks  $T$  and  $E$ . Scores are underlined if their difference is greater than 0.5.

Model	Untrue to Image	Lack of Justification	Non-sensical Sentence
PJ-X	25.0%	26.4%	8.9%
RVT	20.4%	24.2%	12.0%
FME	21.8%	<b>23.1%</b>	13.7%
e-UG	<b>15.9%</b>	25.0%	<b>7.4%</b>
Dataset			
e-SNLI-VE	21.3%	28.7%	12.8%
VCR	21.0%	31.2%	11.7%
VQA-X	20.0%	15.4%	7.4%

Table 5: Main shortcomings of the generated explanations, by models and by datasets. Human judges could choose multiple shortcomings per explanation. The best model results are in bold.

Metric	All datasets	VQA-X	e-SNLI-VE	VCR
BLEU-1	0.222	0.396	0.123	<i>0.032</i>
BLEU-2	0.236	0.412	0.142	<i>0.034</i>
BLEU-3	0.224	0.383	0.139	<i>0.039</i>
BLEU-4	0.216	0.373	0.139	<i>0.038</i>
METEOR	0.288	<b>0.438</b>	0.186	0.113
ROUGE-L	0.238	0.399	0.131	<i>0.050</i>
CIDEr	0.245	0.404	0.133	<i>0.093</i>
SPICE	0.235	0.407	0.162	0.116
BERTScore	<b>0.293</b>	0.431	0.189	<b>0.138</b>
BLEURT [41]	0.248	0.338	<b>0.208</b>	0.128

Table 6: Correlation between human evaluation and automatic NLG metrics on NLEs. All values, except those in *italic*, have p-values  $< 0.001$ .

fit from this additional signal. Indeed, the model is forced to learn a representation of the image and question from which both the answer and explanation can be extracted, which could improve the model’s representation capabilities. To verify this hypothesis, we compare the task scores of modules  $M_T$  that trained only on task  $T$  and those that, together with  $M_E$ , were jointly trained on tasks  $T$  and  $E$ . We do this for e-UG on all three datasets, and for FME and PJ-X on VCR and e-SNLI-VE (because a larger pre-training dataset exists for VQA-X). The results in Table 4 show that, without

any adaptations, the task performance for joint training is equal or better in all but one model-dataset combination. These results suggests that explanations may have the potential to act as “learning instructions” and thereby improve the classification capabilities of a model. Additional experiments are required to further verify this and to develop approaches that more efficiently leverage the explanations.

## 6. Summary and Outlook

We addressed the lack of comparison between existing VL-NLE methods by introducing e-ViL, a unified and reusable benchmark on which we evaluated four different architectures using human judges. We also introduced e-SNLI-VE, the largest existing VL dataset with human-written explanations. The e-ViL benchmark can be used by future works to compare their VL-NLE models against existing ones. Furthermore, our correlation study has shown that automatic NLG metrics have a weak correlation with human judgment. In this work, we also propose a new model, e-UG, which leverages contextualized embeddings of the image-question pairs and achieves a state-of-the-art performance by a large margin on all datasets.

Important questions that need to be addressed in future work are the faithfulness of the explanations (i.e., that they faithfully reflect the model reasoning) and the development of automatic NLG metrics that have a stronger correlation with human judgment.

## Acknowledgements

Maxime Kayser, Leonard Salewski, and Cornelius Emde are supported by Elsevier BV, the International Max Planck Research School for Intelligent Systems, and by Cancer Research UK (grant number C2195/A25014), respectively. This work has been partially funded by the ERC (853489—DEXIM) and by the DFG (2064/1—Project number 390727645). This work has also been supported by the Alan Turing Institute under the EPSRC grant EP/N510129/1, by the AXA Research Fund, the ESRC grant “Unlocking the Potential of AI for English Law”, the EPSRC grant EP/R013667/1, and by the EU TAILOR grant. We also acknowledge the use of Oxford’s Advanced Research Computing (ARC) facility, of the EPSRC-funded Tier 2 facility JADE (EP/P020275/1), and of GPU computing support by Scan Computers International Ltd.



## References

- [1] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.
- [2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic propositional image caption evaluation. In *Proceedings of the European Conference on Computer Vision*. Springer, 2016.
- [3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2015.
- [5] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bénéttot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58, 2020.
- [6] Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, July 2020. Association for Computational Linguistics.
- [7] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.
- [8] Satantjeet Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005.
- [9] Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen tau Yih, and Yejin Choi. Abductive common-sense reasoning. In *International Conference on Learning Representations*, 2020.
- [10] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2015.
- [11] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-SNLI: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.
- [12] Oana-Maria Camburu, Brendan Shillingford, Pasquale Minervini, Thomas Lukasiewicz, and Phil Blunsom. Make up your mind! Adversarial generation of inconsistent natural language explanations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, July 2020.
- [13] Oana-Maria Camburu, Eleonora Giunchiglia, Jakob Foerster, Thomas Lukasiewicz, and Phil Blunsom. The struggles of feature-based explanations: Shapley values vs. minimal sufficient subsets. In *AAAI Workshop on Explainable Agency in Artificial Intelligence*, 2021.
- [14] Rich Caruana. Multitask learning. *Machine Learning*, 28(1), 1997.
- [15] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *Proceedings of the European Conference on Computer Vision*. Springer, 2020.
- [16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019.
- [17] Radhika Dua, Sai Srinivas Kancheti, and Vineeth N. Balasubramanian. Beyond VQA: Generating multi-word answer and rationale to visual questions. *arXiv:2010.12852*, October 2020.
- [18] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.
- [19] Albert Gatt and Emiel Krahmer. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61, 2018.
- [20] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*, 2018.
- [21] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 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*, 2017.

- [22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- [23] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. Generating visual explanations. In *Proceedings of the European Conference on Computer Vision*. Springer, 2016.
- [24] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. Interpreting interpretability: Understanding data scientists’ use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020.
- [25] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the International Conference on Machine Learning (ICML)*, 2018.
- [26] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual explanations for self-driving vehicles. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.
- [27] Neema Kotonya and Francesca Toni. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, November 2020. Association for Computational Linguistics.
- [28] Sawan Kumar and Partha Talukdar. NILE: Natural language inference with faithful natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8730–8742, Online, July 2020. Association for Computational Linguistics.
- [29] Qing Li, Qingyi Tao, Shafiq Joty, Jianfei Cai, and Jiebo Luo. VQA-E: Explaining, elaborating, and enhancing your answers for visual questions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.
- [30] Chin-Yew Lin and Franz Josef Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, 2004.
- [31] CY LIN. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, Barcelona, Spain*, pages 74–81, 2004.
- [32] Shikun Liu, Edward Johns, and Andrew J Davison. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.
- [33] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- [34] Ana Marasović, Chandra Bhagavatula, Jae sung Park, Ronan Le Bras, Noah A Smith, and Yejin Choi. Natural language rationales with full-stack visual reasoning: From pixels to semantic frames to commonsense graphs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, 2020.
- [35] Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. WT5?! Training Text-to-Text Models to Explain their Predictions. *arXiv:2004.14546 [cs]*, April 2020.
- [36] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002.
- [37] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [38] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8), 2019.
- [39] Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. Explain yourself! Leveraging language models for commonsense reasoning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.
- [40] Shaoqing Ren, Kaiming He, Ross B Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2015.
- [41] Thibault Sellam, Dipanjan Das, and Ankur Parikh. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.
- [42] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.
- [43] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.
- [44] Adina Williams, Nikita Nangia, and Samuel Bowman. 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)*, 2018.

- [45] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online, October 2020. Association for Computational Linguistics.
- [46] Jialin Wu and Raymond Mooney. Faithful multimodal explanation for visual question answering. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2019.
- [47] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaying Zhang, Yuxin Peng, and Zheng Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.
- [48] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, January 2019.
- [49] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2, 2014.
- [50] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.
- [51] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTscore: Evaluating text generation with BERT. In *Proceedings of the International Conference on Learning Representations*, 2020.