S^3C: Semi-Supervised VQA Natural Language Explanation via Self-Critical Learning

Wei Suo[1*], Mengyang Sun[2*], Weisong Liu[1], Yiqi Gao[1], Peng Wang[1†], Yanning Zhang[1†], Qi Wu[3]
[1]School of Computer Science and Ningbo Institute, Northwestern Polytechnical University, China.
[2]School of Cybersecurity, Northwestern Polytechnical University, China.
[3]University of Adelaide, Australia.
{suowei1994,sunmenmian,liuweisong,gyqjz}@mail.nwpu.edu.cn
{peng.wang,ynzhang}@nwpu.edu.cn, qi.wu01@adelaide.edu.au

Abstract

VQA Natural Language Explanation (VQA-NLE) task aims to explain the decision-making process of VQA models in natural language. Unlike traditional attention or gradient analysis, free-text rationales can be easier to understand and gain users' trust. Existing methods mostly use post-hoc or self-rationalization models to obtain a plausible explanation. However, these frameworks are bottlenecked by the following challenges: 1) the reasoning process cannot be faithfully responded to and suffer from the problem of logical inconsistency. 2) Human-annotated explanations are expensive and time-consuming to collect. In this paper, we propose a new Semi-Supervised VQA-NLE via Self-Critical Learning (S^3C), which evaluates the candidate explanations by answering rewards to improve the logical consistency between answers and rationales. With a semi-supervised learning framework, the S^3C can benefit from a tremendous amount of samples without human-annotated explanations. A large number of automatic measures and human evaluations all show the effectiveness of our method. Meanwhile, the framework achieves a new state-of-the-art performance on the two VQA-NLE datasets.

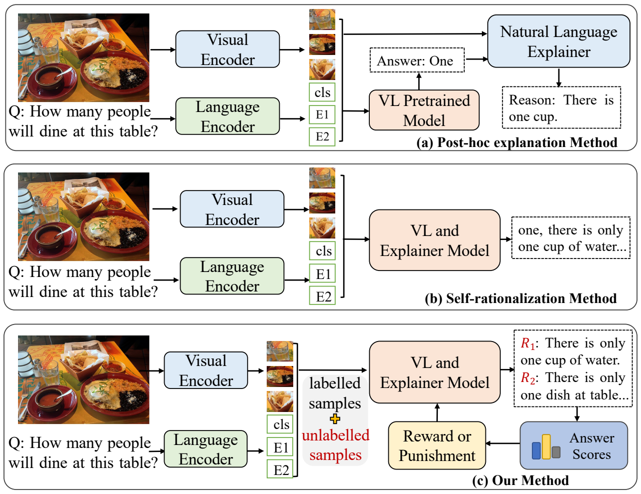

Q: How many people will dine at this table?
Visual Encoder
Language Encoder
cls
E1
E2
Answer: One
VL Pretrained Model
Natural Language Explainer
Reason: There is one cup.
(a) Post-hoc explanation Method
VL and Explainer Model
one, there is only one cup of water...
(b) Self-rationalization Method
labelled samples
unlabelled samples
R_1: There is only one cup of water.
R_2: There is only one dish at table...
Reward or Punishment
Answer Scores
(c) Our Method

Figure 1. Paradigm comparison of different VQA-NLE methods. (a) Post-hoc explanation method adopts two independent models to predict answers and explanations respectively. (b) Self-rationalization method uses a united VL model to simultaneously generate answers and explanations. (c) Our self-critical strategy utilizes answer scores as rewards and obtains more reliable rationales with semi-supervised learning.

1. Introduction

Deep neural networks have enabled significant breakthroughs in a variety of vision-language (VL) tasks such as image captioning [10, 47] and visual question answering (VQA) [2, 39]. Unfortunately, most of them are black box systems, which makes it challenging to gain users' trust [20]. Explaining the decision-making process of deep VL models is a long-standing and essential problem. Some approaches depend on attention mechanisms [2, 30] or gradient-based localization [50] to acquire visual explanations, which can highlight some contributing image regions for the predicted answers. However, simple visualization cannot explain how these areas support the answers and they are also hard to comprehend [20, 48]. Conversely, Natural Language Explanation (NLE) task [6, 38] can explain the decision-making process of a model by generating a natural language sentence. The language-based explanations are more accessible for users to understand, and they can also help researchers optimize the structure of models [34].

Recently, some models of NLE in the VL commu-

[*]These authors contributed equally to this work.
[†]Corresponding authors.

nity have achieved pretty-well results, especially for VQA-NLE [20,34,41,48,58]. They can guide models to generate natural language sentences and interpret how the models get answers. Specifically, the first research line usually treats VQA-NLE as a *predict-then-explain* task [20, 34, 41, 58], namely post-hoc explanations method. As shown in Fig. 1 (a), these methods first depend on pre-trained VL models (such as UNITER [8] or Oscar [25]) to gain answers. Then the fused multi-modal features and the predicted answers are fed into a separated language model (*e.g.,* LSTM [16] or Transformer [54]) to generate corresponding explanations. As shown in Fig. 1 (b), the other line [48] relies on a united VL model while generating both answers and explanations, which is known as the self-rationalization method. This framework can simultaneously predict an answer and generate a rationale by formulating the answer as a text-generation task along with the explanation.

Though significant progress has been made, the two paradigms are still restricted by the following challenges: 1) For the first paradigm, since the decision-making model and interpretation part are two separate modules, it would inevitably lead to unfaithful responses to the reasoning process of the decision models. 2) Due to the lack of explicitly logical relationship modeling, previous work [19] has proved that the straightforward self-rationalization frameworks suffer from the problem of logical inconsistency. 3) The above strategies all require an amount of human-annotated explanations, which are expensive and time-consuming to collect [62].

To solve the above challenges, inspired by [5, 51], we argue that a reasonable rationale can assist the model in obtaining a correct answer, and vice versa, the answer can be converted as an evaluation criterion for possible explanations. In this paper, we propose a new **S**emi-**S**upervised VQA-NLE method with **S**elf-**C**ritical learning, which is called S^3C for short. As shown in Fig.1 (c), given images and related questions, we first leverage a prompting mechanism to construct answer and explanation templates, which can guide the pre-trained VL model to generate answers and multiple candidate explanations based on sequence sampling [2]. Then we design a new self-critical method that converts the answer scores as rewards and encourages the model to generate the explanations which contribute to improving the answer scores. In particular, to reduce the dependency on expensive human annotations, we further extend our method to the semi-supervised version, which utilizes the unlabelled samples [1] (*i.e.*, conventional VQA data [4, 36]) to significantly enhance the self-interpretability of the model. With the self-critical strategy and the semi-supervised learning, our method effectively models the logical relationships and promotes the logical consistency between answer-explanation pairs. According to automatic measures and human evaluations, the S^3C outperforms the state-of-the-art models for the VQA-NLE task on the widely used two datasets and provides a new paradigm for our community. In summary, we make the following contributions:

1) We propose a new self-critical VQA-NLE method that can model the logical relationships between answer-explanation pairs and evaluate the generated rationales by answering rewards. This strategy effectively improves the logical consistency and the reliability of the interpretations.

2) We develop an advanced semi-supervised learning framework for VQA-NLE, which utilizes amounts of samples without human-annotated explanations to boost the self-interpretability of the model further. To the best of our knowledge, we are the first to explore semi-supervised learning on the VQA Natural Language Explanation.

3) The proposed S^3C achieves new state-of-the-art performance on VQA-X [13] and A-OKVQA [49] benchmark datasets. Meanwhile, automatic measures and human evaluations all show the effectiveness of our method.

2. Related work

2.1. Explainability in Visual Question Answering

The visual question answering (VQA) is firstly proposed by [33] that requires an intelligent agent to generate an answer by giving an image and a question. Many approaches have been introduced such as joint embedding [13, 61], attention mechanisms [3, 31], memory networks [32, 59] and graph neural networks [23, 56]. Although the VQA task has been well studied, the reasoning process of the models is always agnostic. Some methods apply visualization technologies to achieve visual explanation, such as Grad-CAM [50] and U-CAM [42]. However, because image visualization cannot support the answer based on the attended areas [57], in this paper, we focus on improving free-text explanations that are more convenient and easier for users to understand. In this topic, the early work is proposed by [41]. The paper conducts the VQA-X dataset and utilizes human annotations to inspire the decision-making process of VQA models. [20] designs a new model that combines a pre-trained language model and a VL model to generate free-text explanations. [60] combines stronger pre-trained VL model (*i.e.* Oscar [25]) and generation model (*i.e.* GPT-2 [46]) to obtain better results. Recently, [48] proposes a unified model which can simultaneously predict answers and explanations based on a pre-trained caption model. Unlike previous methods, we introduce a new self-critical strategy to model the logical relationships between answers and rationales. It can encourage the model to enhance logical consistency and generate more reasonable explanations.

[1] In this paper, we use “unlabelled samples” and “labelled samples” to indicate the question-answer (QA) pairs without/with human explanations.

2.2. Pre-trained models and Prompt learning

Pre-trained models have been applied in many fields, such as various NLP tasks [12,46] and VL tasks [29,52,65]. Most of these pre-trained models utilize a stack of Transformer structures as the backbone. To generalize the pre-trained models to other downstream tasks, previous works mostly fine-tune whole models on each downstream VL task. However, the ability of pre-trained models would be limited due to the mismatch between pre-trained tasks and downstream tasks. Hence, prompt learning [9, 18, 45, 64] is proposed, and it can keep the optimization consistency. [45] designs the templates to transfer the knowledge to downstream tasks. [9] uses a textual generation framework for uniform optimization. In this paper, we use the pre-trained model based on the image caption task as our backbone. It has been proved that these additional pre-train tasks (*e.g.,* image feature regression or mask language tokens) cannot significantly improve the performance for explanations [48]. Meanwhile, we apply different language prompt templates to motivate the model to generate corresponding answers and rationales.

2.3. Semi-supervised learning

The development of deep learning depends on a large number of labelled data, while there are many cases in that only a small amount of data can be obtained [44]. To solve this challenge, Semi-supervised learning is proposed [44], which aims to train models using a small number of labelled data and amounts of unlabelled data. For example, [22] proposes a generatively semi-supervised framework based on variational autoencoders, and it can jointly optimize the model and variational parameters. Recently, [63] simultaneously leverages self-supervised and semi-supervised learning to address image classification tasks. To our knowledge, our work is the first semi-supervised learning framework for the VQA-NLE task, which effectively alleviates the reliance on expensive human annotations and further boost the self-interpretability of the model.

3. Method

In this section, we introduce our Semi-Supervised VQA-NLE method with the Self-Critical learning (S^3C) framework. Our aim is to strengthen the logical consistency between answer-explanation pairs and improve the reliability of the rationales. As shown in Fig. 2, the S^3C comprises an "Answer-Explanation Prompt" module and a "Self-Critical Reinforcement" module. Unlike previous approaches, our method first uses a prompting mechanism to generate answers and candidate explanations. Then, we design a new self-critical module that converts the answer scores as rewards to evaluate these reasons. Furthermore, this strategy can conveniently apply the unlabelled QA pairs to enrich training data and enhance the self-interpretability. Next, we describe the components of our model in detail.

3.1. Pre-trained Vision-Language Backbone

Given an image $I \in \mathbb{R}^{W\times H\times 3}$ and a natural language question $Q = \{q_t\}_{t=1}^{T}$, where q_t represents the t-th word, T is the length of the question and $W \times H \times 3$ denotes the size of the image. Our goal is to predict the answer and generate a corresponding free-text rationale. Following previous works [48, 49], we adopt the CLIP vision encoder [45] and a pre-trained image caption model (*i.e.*, ClipCap [37]) as our basic backbone. During pre-training, the VL model uses image embeddings from the CLIP as prefixes and fine-tunes a language model (*i.e.,* GPT-2 [46]) to generate image captions. We refer readers to [37] for more information about the pre-trained model. In this paper, we consider the image I and question Q as the prefixes of the answer-explanation sequences. Specifically, we first apply ViT-B [14] and "classification token" from the CLIP to obtain the image features. Then, a group of light and simple Multi-Layer Perceptron (MLP) is utilized for transforming the image features to the $V = \{v_s\}_{s=1}^{S}$, $v_s \in \mathbb{R}^c$, where the dimension size $c = 768$ and the image sequence length $S = 10$. The above computations can be formulated as follows:

$$v_1, v_2, \cdots, v_S = \mathrm{MLP}(\mathrm{CLIP}(I)). \tag{1}$$

Note that we only update the mapping network (*i.e.*, MLP) during training, while the original visual encoder parameters from the CLIP would remain frozen. For question Q, each word q_t would be mapped to the corresponding word embedding $e_t \in \mathbb{R}^c$ by the pre-trained caption model. Finally, we obtain the image and question sequences Z:

$$Z = [\overbrace{v_1, v_2, \cdots, v_S}^{\text{image embedding}}, \underbrace{e_1, e_2, \cdots, e_T}_{\text{question embedding}}], \tag{2}$$

where Z is the concatenated multi-modal prefixes.

3.2. Answer-Explanation Prompt Module

It has been proved that the prompting mechanism can maintain the same optimization objectives between pre-trained and downstream tasks [18, 28]. Considering the convenience and explainability, we leverage hand-crafted prompts as templates to enable the model to generate answers or explanations. As shown in Fig. 2, the S^3C includes two different kinds of templates, which are introduced respectively as follows.

Base Answer Template. The core idea of our S^3C is to use the answer scores as evaluation criteria. Hence, we first establish a basic answer template to acquire base answer scores. Different from previous works [20, 41], our

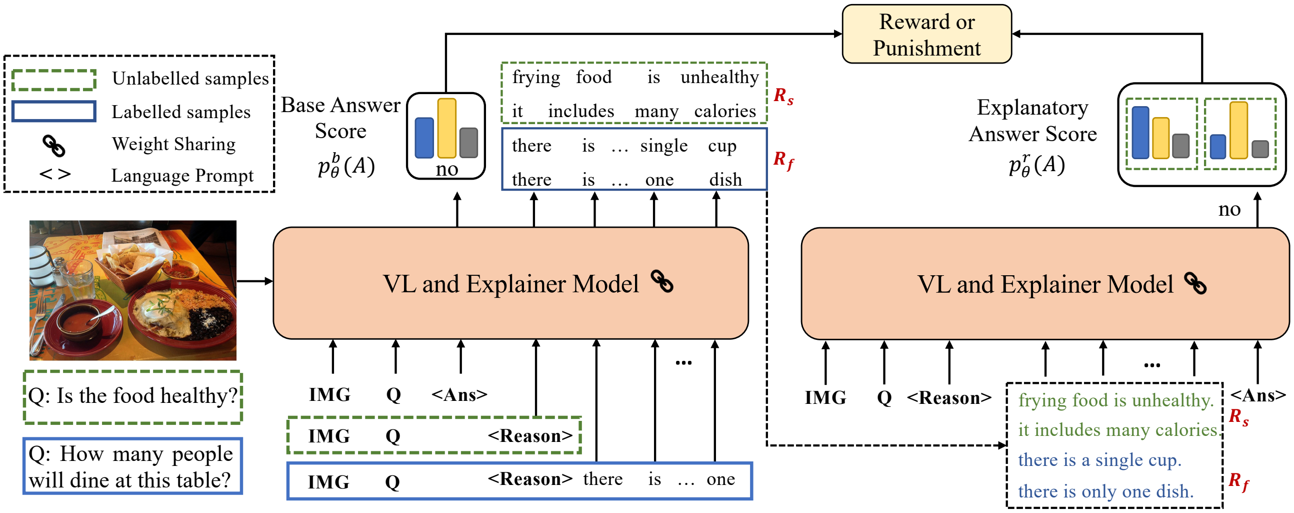


Figure 2. Overview of our Semi-Supervised VQA-NLE method via Self-Critical learning (S^3C) framework. Given images and corresponding questions for labelled and unlabelled samples, we first use Answer-Explanation Prompt module to obtain the base answer scores and candidate explanations with a pre-trained VL model. Then these reasons are reorganized and fed back into the model to capture the explanatory answer score. Further, our Self-Critical Reinforcement module evaluates the generated explanations and returns the rewards to improve the self-interpretability of the model.

method treats the visual question answering task as a generation task so that the model can produce answers without a predefined answer space. Specifically, for a given image and question sequence Z, we use natural language tokens "the answer is" to inspire the model to generate proper answers. By concatenating the language prompt, the base answer template $Z_a = [Z; \langle answer\rangle]$ can be obtained, where $[;]$ and $\langle answer\rangle$ indicate concatenation operation and the specific language prompts, respectively. During training, the base answer template and ground-true answer label $A = \{a_n\}_{n=1}^N$ are fed into the VL model, where a_n is the n-th answer token and N denotes the length of the answer label. Next, we compute the answer loss conditioned on the Z_a in an autoregressive fashion:

$$L_a = -\frac{1}{N}\sum_{n=1}^{N} logP(a_n|Z_a, a_1, a_2, \cdots, a_{n-1}; \theta), \quad (3)$$

where θ denotes the parameters of the VL model. Moreover, based on the indexes of ground-true answers, we can acquire average probability $p_\theta^b(A)$ as our base answer score.

Explanation Generation Template. To produce a reasonable rationale, we leverage the language prompt "the reason is" to motivate the model to generate free-text explanations. Like the base answer template, an explanation generation template $Z_e = [Z; \langle reason\rangle]$ is constructed, where $\langle reason\rangle$ is the natural language tokens. For labelled samples, we follow Eq. 3 in the autoregressive fashion and the cross-entropy loss to compute explanation loss L_e. Because there are no human-annotated explanations available, we would not compute any loss for unlabelled QA samples.

3.3. Self-Critical Reinforcement Module

To gain logically consistent rationales, in this module, we expand the searching space by introducing sequence sampling algorithm [2] and generate a set of candidate explanations. Besides, the answer scores are treated as rewards to encourage the model to output more detailed interpretations. It's worth noting that the above operations would be implemented on both labelled and unlabelled samples.

Candidate Explanation Generation. Benefiting from the language-based prompting strategy and the pre-trained VL model, our explanation generation template can easily guide the model to produce human-readable sentences. Hence, for unlabelled QA samples, we directly apply the explanation generation template Z_e to generate the candidate rationales. To be more specific, we utilize beam search [2] to sample the top-K words from the VL model probability distribution at each time step and maintain these sequences with the highest probability. Then, these generated sentences are integrated into candidate explanations $R_s = \{r_k^s\}_{k=1}^K$ for each QA pair without human-annotated explanations, where r_k^s and K indicate k-th rationale and the size of beam search, respectively. Moreover, for labelled samples, a similar mechanism is used to build the corresponding explanation set $R_f = \{r_k^f\}_{k=1}^K$. The above operations are implemented in all samples for the following reasons: 1) Larger search space. As shown in Fig. 2, "frying food is unhealthy" and "it includes many calories" can both be rational explanations for the answer “no”. The expanded search space provides more possibilities for generating reliable rationales. 2) Avoid overfitting. For labelled samples, although we can depend on human explanations to train the model, these labels are still one-sided and subjective. Using the sequence sampling strategy can prevent the model from overfitting these specific annotations [2,47]. In the end, because the labelled and unlabelled samples would be integrated into a mini-batch during training, we simplify R_f and R_s into $R = \{r_k\}_{k=1}^{K}$ to indicate the candidate explanations for each sample.

Self-Critical Reward. Inspired by [5,51], an ideal rationale can help the model to infer the answer better. Based on this insight, we argue that the answer scores can be converted as self-critical rewards to evaluate these candidate explanations. Considering that this is a non-differentiable operation, we adopt reinforcement learning method [47] to achieve end-to-end training. In particular, given a sample and corresponding candidate explanations R, we design a new input template:

$$Z_r^k = [Z; \langle reason \rangle; r_k; \langle answer \rangle], \tag{4}$$

where Z_r^k denotes the template of adding possible rationale r_k. Then, this template is fed back into the model and obtains average probability $p_\theta^r(A)$ about the answer, namely explanatory answer scores, where A is the ground-truth answer for the sample. Meanwhile, we use the average probability $p_\theta^b(A)$ from the output of the base answer template as the base scores. By applying reinforcement learning, the gradient is calculated by:

$$\nabla_\theta L_r(\theta) = -\frac{1}{K}\sum_{k=1}^{K}(p_\theta^r(A) - p_\theta^b(A))\nabla_\theta log p_\theta(r_k), \tag{5}$$

where $p_\theta(r_k)$ is the probability of k-th explanations. Based on the above computation, this gradient would tend to increase the probability of k-th rationales when the answer score $p_\theta^r(A)$ higher than the scores $p_\theta^b(A)$ from the base answer template. Finally, for labelled samples, we append the human-annotated rationales to the candidate explanation R and predict the answers by cross-entropy loss, namely L_{ea}.

3.4. Loss

During the process of training, the overall loss function can be represented as follows:

$$L = L_a + L_e + L_{ea} + \lambda L_r, \tag{6}$$

where λ is used for balancing these two different types of losses (*i.e.,* cross-entropy loss and reinforcement loss). At inference time, our model would first generate the rationales about QA pairs, then we use the explanatory template (*i.e.*, Eq. 4) to obtain the corresponding answers.

4. Experiment

4.1. Experimental setting

Datasets. Following previous methods [48, 49], we mainly carry out the experiments on the two different VQA-NLE datasets: VQA-X [41] and A-OKVQA [49]. Meanwhile, since explain annotations are expensive and time-consuming, we also utilize large-scale VQA v2.0 [15] and OK-VQA [36] datasets to build the semi-supervised learning paradigm. Next, we will introduce the four datasets.

VQA-X. It is a vision-language dataset that provides explanations for justifying the answers. VQA-X is collected from the Visual Question Answering (VQA) dataset [4] where the images are obtained from the MSCOCO [27]. It consists of 28K images and 33K QA pairs, split into 29K/1.4K/1.9K for training, validation and testing. Meanwhile, VQA-X constructs complementary pairs which provide a question and two semantically similar images with different answers.

A-OKVQA. Compared to the VQA-X, the questions of A-OKVQA generally are required commonsense reasoning about the scene described in the images. It includes 24,903 Question/Answer/Rationale triplets, split into 17.1K/1.1K/6.7K for training, validation and testing. It collects images from the COCO 2017 [7] dataset, and is further filtered to obtain 23.7K unique images. Compared with previous datasets, the A-OKVQA has richer questions and requires broader areas of knowledge for reasoning.

VQA & OK-VQA. We use VQA v2.0 and OK-VQA to provide large-scale unlabelled datasets for semi-supervised learning. The VQA v2.0 dataset is widely used for many previous works [2, 52]. It consists of 443k questions and 195k images. To select explainable questions instead of some obvious cases (*e.g.*, How many...?, What color...?), we filter out these questions from VQA v2.0 and obtain the $\sim$90k additional questions based on the rules [41]. On the other hand, we also use knowledge-based OK-VQA [36] dataset to provide unlabelled knowledge-based QA pairs. It contains a total of 14k questions on 14k images.

Implementation Details. Following [37], each image is first pre-processed (such as including image resize, center crop, and normalize) by CLIP [45]. Meanwhile, we fix the ViT-B weights from the CLIP visual encoder to accelerate the training speed. For the mapping network, the image sequence length $S = 10$ and the embedding size is 768. The AdamW [21] is used as our optimizer with the weight decay 1e-5, and the batch size and beam size K are set to 4 and 2. The weight coefficient λ is set to 10. We train all models on the four 1080Ti GPUs for 30 epochs with a learning rate of 1e-5.

4.2. Evaluation Measures

Automatic Metering. Following [48,60], we use the auto-

Table 1. Comparison with the state-of-the-art methods on the **VQA-X**. Note that these results are **unfiltered** scores. S^3C^* denotes the model without unlabelled samples.

	VQA-X						
Approach	B4	M	R	S	C	Acc	Human
CAPS [41]	5.9	12.6	26.3	11.9	35.2	68.6	-
PJ-X [41]	19.5	18.2	43.4	15.1	71.3	76.4	65.4
FME [58]	24.4	19.5	47.7	17.9	88.8	75.5	-
NLX-GPT [48]	25.6	21.5	48.7	20.2	97.2	83.1	70.2
S^3C^*(ours)	26.5	22.0	49.0	20.9	100.5	83.7	73.9
S^3C (ours)	**27.8**	**22.8**	**50.7**	**21.5**	**104.4**	**85.6**	**77.4**

Table 2. Comparison with the state-of-the-art methods on the **A-OKVQA**. Note that these results are **unfiltered** scores. S^3C^* denotes the model without unlabelled samples.

	AOKVQA							
Approach	B4	M	R	S	C	Acc val	Acc test	Human
ViLBERT [29]	-	-	-	-	-	30.6	25.9	-
LXMERT [52]	-	-	-	-	-	30.7	25.9	-
KRISP [35]	-	-	-	-	-	33.7	27.1	-
Clipcap [49]	-	-	-	-	-	30.8	25.9	-
e-UG [20]	15.1	18.1	42.4	14.9	51.5	30.5	25.6	44.1
NLX-GPT [48]	20.1	17.0	46.3	15.8	65.4	32.7	28.7	46.9
S^3C^*(ours)	21.8	17.9	47.3	17.3	70.6	33.0	29.6	49.4
S^3C (ours)	**22.5**	**18.5**	**48.4**	**18.1**	**74.4**	**34.2**	**33.5**	**54.7**

Table 3. Comparison with the state-of-the-art methods on the VQA-X. Note that these results are **filtered** scores. S^3C^* denotes the model without unlabelled samples.

	B1	B2	B3	B4	M	R	S	C	Acc	Human
RVT [34]	51.9	37.0	25.6	17.4	19.2	42.1	15.8	52.5	68.6	60.5
PJ-X [41]	57.4	42.2	30.9	22.7	19.7	46.0	17.1	82.7	76.4	69.3
FME [58]	59.1	43.4	31.7	23.1	20.4	47.1	18.4	87.0	75.5	-
QA-only [20]	51.0	36.4	25.3	17.3	18.6	41.9	14.9	49.9	-	-
e-UG [20]	57.3	42.7	31.4	23.2	22.1	45.7	20.1	74.1	80.5	71.4
NLX-GPT [48]	64.2	49.5	37.6	28.5	23.1	51.5	22.1	110.6	83.2	73.7
S^3C^*(ours)	64.4	49.9	38.0	29.1	23.4	51.9	22.7	112.1	83.7	75.9
S^3C (ours)	**64.7**	**50.5**	**38.8**	**30.7**	**23.9**	**52.1**	**23.0**	**116.7**	**85.6**	**79.2**

matic metrics BLEU [40], METEOR [11], ROUGE-L [26], SPICE [1] and CIDEr [55] to evaluate generated explanations. For the evaluation of predicted answers, we follow [48, 60] to compute the VQA accuracy.

Human Evaluation. Automatic VQA-NLE measures do not always reflect the correctness and logicality of the explanations [20, 53], thus we also build human evaluations. The process is similar to [20, 34]. Specifically, for each explanation, three human evaluators are required to decide whether an explanation can justify the answer and select an option (including "yes, weak yes, weak no and no"). The selection will be mapped to scores (1, $\frac{2}{3}$, $\frac{1}{3}$ and 0). The final scores are computed by averaging among all test samples. Meanwhile, these evaluators are asked to choose the reasons for unqualified explanations. We follow [20] to define three kinds of aspects to evaluate the explanations: irrelevant explanations, insufficient explanations and meaningless explanations. First, the irrelevant explanations may not match the image, for example, "have a long neck" is a good explanation for the answer "giraffes" when asked "what animals are these?", but the image may display cows. Second, the insufficient explanations only describe the image but cannot corroborate the answers. For example, the sentence "there are some people" does not sufficiently justify the question "do people have a party?". Lastly, some nonsensical sentences could be judged as contradictory explanations, such as "a man is a woman". For each sample, the human evaluators can select multiple shortcomings. More details of the human evaluation can be found in the [20].

4.3. Quantitative evaluation.

Automatic Evaluation. We compare our method with the state-of-the-art models on the VQA-X and A-OKVQA datasets in Table 1-3. The B4, M, R, S, C, Acc and Human are short for BLEU-4, METEOR, ROUGE-L, SPICE, CIDEr, Answer precision and Human evaluation. We use "unfiltered" to indicate that the explanations are evaluated regardless of whether the answer is true or false. While "filtered" is to only consider the explanations which have correct answers. From Table 1, the "S^3C*" denotes that we only apply the self-critical framework without using unlabelled samples, while the row "S^3C" indicates our method in the semi-supervised setting. We observe that the proposed framework outperforms both the post-hoc explanation methods [41, 58] and the self-rationalization method [48]. Meanwhile, it's worth noting that although the NLX-GPT uses a more powerful pre-trained VL model with larger pre-trained datasets (*i.e.*, COCO captions [27], Flicker30k [43] and Visual Genome [24]), our method still obtains 3.3 absolute gains on the CIDEr indicator with fewer pre-trained data (only using COCO captions). Furthermore, when we utilize our proposed semi-supervised paradigm, the results are further improved by 7.2 points. These results show that our model can generate more reliable explanations and it can benefit from the amount of data without human-explanation labels. We also report the accuracy of answers in the column of "Acc". It can be observed that our self-critical method with semi-supervised learning can simultaneously boost the precision of answers and cor-

Table 4. **Main shortcomings.** The main shortcomings of unqualified explanations on the VQA-X dataset. For each sample, human evaluators can select multiple shortcomings.

Model	Irrelevant explanations	Insufficient explanations	Meaningless explanations
RVT [34]	25.7%	33.5%	11.4%
PJ-X [41]	21.1%	28.4%	9.2%
e-UG [20]	22.8%	25.4%	8.7%
NLX-GPT [48]	20.3%	22.2%	9.1%
S^3C (ours)	**17.3%**	**18.9%**	**8.2%**

Table 5. **Cross-dataset testing.** We alternately use the VQA-X and A-OKVQA as source dataset and target dataset to test the generalization of our framework.

	VQA-X→A-OKVQA					
Approach	B4	M	R	S	C	Acc
NLX-GPT [48]	10.7	12.7	34.2	10.7	35.4	10.4
S^3C (ours)	**12.0**	**13.3**	**34.3**	**12.5**	**45.3**	**18.8**
	A-OKVQA→VQA-X					
Approach	B4	M	R	S	C	Acc
NLX-GPT [48]	9.1	13.6	32.8	9.1	33.2	42.4
S^3C (ours)	**10.9**	**15.0**	**34.1**	**10.4**	**38.6**	**43.8**

Table 6. **Ablation study.** We ablate key components to demonstrate the effectiveness of our method. SCR and Semi are Self-Critical Reinforcement module and Semi-supervised learning paradigm respectively.

	question	image	answer	explanation	SCR	Semi	B4	M	R	S	C	Acc
1	✓	✓	✓	–	–	–	–	–	–	–	–	80.1
2	✓	✓	–	✓	–	–	24.4	20.7	47.3	19.5	90.4	–
3	✓	✓	✓	✓	–	–	27.5	22.9	50.4	21.9	109.1	82.2
4	✓	✓	✓	✓	✓	–	29.1	23.4	51.9	22.7	112.1	83.7
5	✓	✓	✓	✓	✓	✓	**30.7**	**23.9**	**52.1**	**23.0**	**116.7**	**85.6**

responding explanations.

In Table 2, we evaluate our method on the A-OKVQA dataset. The results show that the S^3C can outperform all the previous works [17, 29, 35, 48, 49, 52]. Especially, compared to the SOTA model [48], there are 9.0 points and 4.8% improvements in the CIDEr and answer accuracy based on our model. It demonstrates that our model can benefit from semi-supervised learning and generate explanations about commonsense reasoning.

In addition, to prove the algorithm's validity, we follow [20, 48] to report the filtered scores for the VQA-X dataset in Table 3. Through filtering the correct answers, our method can outperform whatever post-hoc [20, 34, 41, 58] and self-rationalization [48] methods. Meanwhile, the proposed S^3C achieves a new state-of-the-art with the CIDEr score improved by 6.1 points to 116.7 and boost the answer accuracy by 2.4% to 85.6.

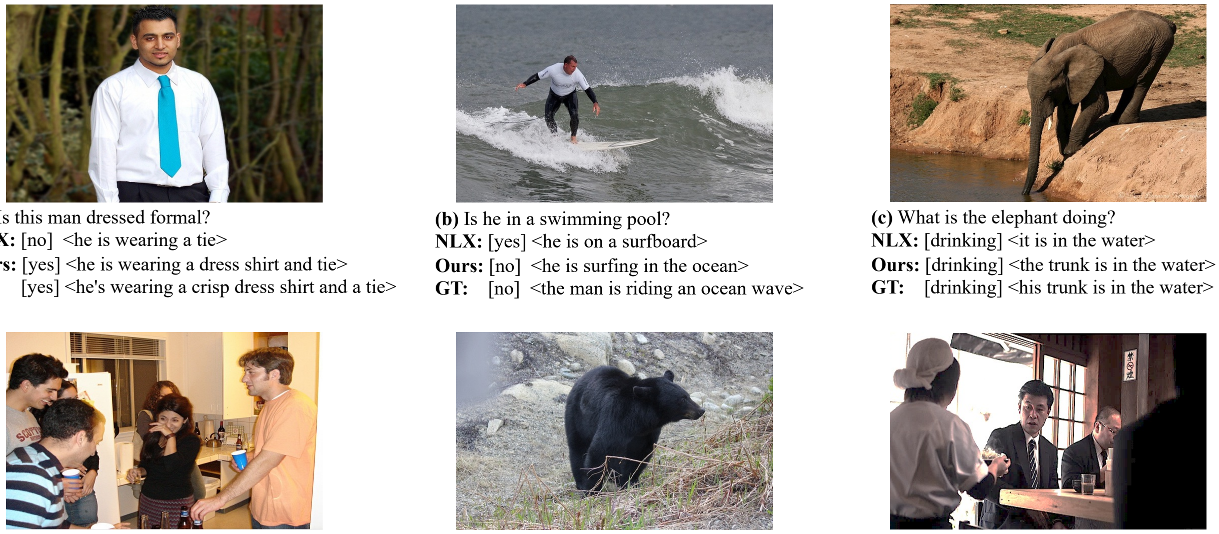
Human Evaluation. To evaluate the faithfulness and correctness of these generated explanations, we conduct the human evaluation that is shown in the column of "Human" in Table 1-3. The experimental results further prove our method has better self-interpretability for the VQA-NLE task. Moreover, we also ask the human evaluators to select the shortcomings for each unqualified explanation on the VQA-X dataset. As shown in Table 4, we build three shortcoming options (*i.e.,* irrelevant explanations, insufficient explanations and meaningless explanations) with a multi-choice process. These results indicate that the S^3C can obtain relatively better rationales and empirically confirm the effectiveness of our method.

4.4. Cross-dataset Performance.

We propose cross-dataset experiments to measure the generalization and reliability of our framework. For all we know, there is very little research on cross-dataset explanation generation. As shown in the first two rows of Table 5, we use the VQA-X and the A-OKVQA as the source dataset and target dataset to test the generalization ability of our framework. Specifically, we use the weights from the state-of-the-art method [48], which are trained on the VQA-X dataset and tested on the A-OKVQA dataset to obtain the results of the NLX-GPT. From the comparisons of six indicators, we find that our self-critical model has better generalization from the general VQA to the knowledge-based VQA. This suggests that S^3C can learn commonsense from visual question answering and generate better explanations and answers. Meanwhile, when we exchange the source dataset and target dataset as shown in the last two rows of Table 5, our framework can still surpass the existing SOTA model on all evaluation measures. These quantitative results effectively demonstrate the generalization and stability of our model.

4.5. Ablation Studies.

As shown in Table 6, we conduct several ablation studies on the VQA-X to demonstrate the effectiveness of our method. In particular, we first build the experiments in the 1-3 rows of Table 6, which use our Answer-Explanation Prompt module with cross-entropy loss as the baseline models. It can be found that when we simultaneously generate both answers and interpretations, the answer accuracy and CIDEr can achieve significant improvements. Meanwhile,



(a) Is this man dressed formal?
NLX: [no] <he is wearing a tie>
Ours: [yes] <he is wearing a dress shirt and tie>
GT: [yes] <he's wearing a crisp dress shirt and a tie>

(b) Is he in a swimming pool?
NLX: [yes] <he is on a surfboard>
Ours: [no] <he is surfing in the ocean>
GT: [no] <the man is riding an ocean wave>

(c) What is the elephant doing?
NLX: [drinking] <it is in the water>
Ours: [drinking] <the trunk is in the water>
GT: [drinking] <his trunk is in the water>

(d) Are the people having a party?
NLX: [no] <there are no people present>
Ours: [yes] <they are laughing and having a good time>
GT: [yes] <they are standing in a kitchen with drinks in hand and laughing>

(e) Does this animal appear to be looking at the camera?
NLX: [yes] <it is looking at something>
Ours: [no] <it is looking away from the camera>
GT: [no] <the animal is looking off to the side>

(f) Is this a bar?
NLX: [yes] <there are people sitting at tables>
Ours: [yes] <the men are sitting at a bar with drinks in front of them>
GT: [yes] <there are drinks on a long countertop with people eating and drinking>

Figure 3. Some generation results on the VQA-X datasets. The [ ] and <> indicate answers and explanations respectively. We show the results of the state-of-the-art NLX-GPT [48], our method and ground-truth (they are short for NLX, Ours and GT).

these results also prove that the answer-explanation pairs have inherent consistency and they can promote each other. Further, we establish the Self-Critical Reinforcement module as shown in the fourth row, and the result of accuracy and CIDEr are boosted by 1.5% and 3.0 points. It suggests that our self-critical method with sequence sampling strategy can encourage the model to generate more reliable explanations and correct answers. Finally, when we extend our model to the semi-supervised setting, the performance improves to 85.6 on the accuracy and 116.7 on the CIDEr over the baseline model (*i.e.,* the row 5 of Table 6) by 3.4% and 7.6 points. The results demonstrate that our framework is remarkably effective in improving the logical consistency and the self-interpretability of the model.

4.6. Qualitative evaluation.

As shown in Fig. 3, we show some qualitative results from NLX-GPT [48] and our S^3C method on the VQA-X dataset. Through overall comparison, our model achieves better logical consistency between answers and explanations. For example, in Fig. 3 (a), although the SOTA method [48] correctly identifies the significant symbol of formal dress (*i.e.*, tie), the predicted answer is "no" that is contradictory to the explanation. On the contrary, our method not only obtains a more complete and faithful rationale but also generates a logically consistent answer. Additionally, the S^3C can also produce more persuasive and trusty sentences. For instance, in Fig. 3 (b)-(f), the generated sentences contain scene description "having a good time" and fine-grained rationale "the trunk is in the water".

5. Conclusion

In this paper, we propose a new Semi-Supervised VQA-NLE method via Self-Critical Learning (S^3C). Different from previous works, our method first utilizes the prompting mechanism to motivate the model to generate answers and candidate explanations. Meanwhile, we design a novel Self-Critical Reinforcement module, which converts the answer scores as rewards to evaluate these possible rationales. Furthermore, our framework can benefit from an abundance of question-answer pairs without human-annotated explanations and further boost the self-interpretability of the model. With the automatic measures and human evaluations, our S^3C achieves a new state-of-the-art on multiple benchmarks and provides a new paradigm for our community.

Acknowledgements

This work was supported by National Key R&D Program of China (No.2020AAA0106900), the National Natural Science Foundation of China (No.U19B2037), Shaanxi Provincial Key RD Program (No.2021KWZ-03), and Natural Science Basic Research Program of Shaanxi (No.2021JCW-03).

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer, 2016.

[2] 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*, pages 6077–6086, 2018.

[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 *CVPR*, pages 6077–6086, 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 IEEE international conference on computer vision*, pages 2425–2433, 2015.

[5] Ermanno Bencivenga. Free logics. In *Handbook of philosophical logic*, pages 147–196. Springer, 2002.

[6] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31, 2018.

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

[8] 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 *European conference on computer vision*, pages 104–120. Springer, 2020.

[9] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR, 2021.

[10] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10578–10587, 2020.

[11] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[13] Jianfeng Dong, Xirong Li, and Cees G M Snoek. Predicting visual features from text for image and video caption retrieval. *IEEE TMM*, 20(12):3377–3388, 2018.

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

[15] 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*, pages 6904–6913, 2017.

[16] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[17] Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Pythia v0. 1: the winning entry to the vqa challenge 2018. *arXiv preprint arXiv:1807.09956*, 2018.

[18] Woojeong Jin, Yu Cheng, Yelong Shen, Weizhu Chen, and Xiang Ren. A good prompt is worth millions of parameters: Low-resource prompt-based learning for vision-language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2763–2775, 2022.

[19] Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. Maieutic prompting: Logically consistent reasoning with recursive explanations. *arXiv preprint arXiv:2205.11822*, 2022.

[20] Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. e-vil: A dataset and benchmark for natural language explanations in vision-language tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1244–1254, 2021.

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[22] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. *Advances in neural information processing systems*, 27, 2014.

[23] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017.

[24] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

[25] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020.

[26] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[28] Yuhang Liu, Wei Wei, Daowan Peng, and Feida Zhu. Declaration-based prompt tuning for visual question answering. *arXiv preprint arXiv:2205.02456*, 2022.

[29] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[30] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. *Advances in neural information processing systems*, 29, 2016.

[31] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *NeurIPS*, pages 289–297, 2016.

[32] Chao Ma, Chunhua Shen, Anthony R. Dick, Qi Wu, Peng Wang, Anton van den Hengel, and Ian D. Reid. Visual question answering with memory-augmented networks. In *CVPR*, pages 6975–6984, 2018.

[33] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. *Advances in neural information processing systems*, 27, 2014.

[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. *arXiv preprint arXiv:2010.07526*, 2020.

[35] Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14111–14121, 2021.

[36] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019.

[37] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.

[38] Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. Wt5?! training text-to-text models to explain their predictions. *arXiv preprint arXiv:2004.14546*, 2020.

[39] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-linear attention networks for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10971–10980, 2020.

[40] 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*, pages 311–318, 2002.

[41] 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*, pages 8779–8788, 2018.

[42] Badri N Patro, Mayank Lunayach, Shivansh Patel, and Vinay P Namboodiri. U-cam: Visual explanation using uncertainty based class activation maps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7444–7453, 2019.

[43] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

[44] Guo-Jun Qi and Jiebo Luo. Small data challenges in big data era: A survey of recent progress on unsupervised and semi-supervised methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

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

[47] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024, 2017.

[48] Fawaz Sammani, Tanmoy Mukherjee, and Nikos Deligiannis. Nlx-gpt: A model for natural language explanations in vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8322–8332, 2022.

[49] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. *arXiv preprint arXiv:2206.01718*, 2022.

[50] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[51] Suzanna Sia, Anton Belyy, Amjad Almahairi, Madian Khabsa, Luke Zettlemoyer, and Lambert Mathias. Logical satisfiability of counterfactuals for faithful explanations in nli. *arXiv preprint arXiv:2205.12469*, 2022.

[52] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[53] Rakesh Vaideeswaran, Feng Gao, ABHINAV MATHUR, and Govind Thattai. Towards reasoning-aware explainable vqa. In *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022*.

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[55] 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*, pages 4566–4575, 2015.

[56] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *ICLR*, 2018.

[57] Jialin Wu and Raymond Mooney. Self-critical reasoning for robust visual question answering. *Advances in Neural Information Processing Systems*, 32, 2019.

[58] Jialin Wu and Raymond J Mooney. Faithful multimodal explanation for visual question answering. *arXiv preprint arXiv:1809.02805*, 2018.

[59] Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. pages 2397–2406, 2016.

[60] Qian Yang, Yunxin Li, Baotian Hu, Lin Ma, Yuxin Ding, and Min Zhang. Chunk-aware alignment and lexical constraint for visual entailment with natural language explanations. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3587–3597, 2022.

[61] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Hierarchy parsing for image captioning. In *IJCV*, pages 2621–2629, 2019.

[62] Yordan Yordanov, Vid Kocijan, Thomas Lukasiewicz, and Oana-Maria Camburu. Few-shot out-of-domain transfer learning of natural language explanations. *arXiv preprint arXiv:2112.06204*, 2021.

[63] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1476–1485, 2019.

[64] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022.

[65] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pretraining for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13041–13049, 2020.