Towards More Faithful Natural Language Explanation Using Multi-Level Contrastive Learning in VQA

Chengen Lai, Shengli Song*, Shiqi Meng, Jingyang Li, Sitong Yan, Guangneng Hu

School of Computer Science and Technology, Xidian University, Xi'an, China
{laice, ShiqiMeng, jylee, styan}@stu.xidian.edu.cn, shlsong@xidian.edu.cn, njuhgn@gmail.com

Abstract

Natural language explanation in visual question answer (VQA-NLE) aims to explain the decision-making process of models by generating natural language sentences to increase users' trust in the black-box systems. Existing post-hoc methods have achieved significant progress in obtaining a plausible explanation. However, such post-hoc explanations are not always aligned with human logical inference, suffering from the issues on: 1) Deductive unsatisfiability, the generated explanations do not logically lead to the answer; 2) Factual inconsistency, the model falsifies its counterfactual explanation for answers without considering the facts in images; and 3) Semantic perturbation insensitivity, the model can not recognize the semantic changes caused by small perturbations. These problems reduce the faithfulness of explanations generated by models. To address the above issues, we propose a novel self-supervised **M**ulti-level **C**ontrastive **L**earning based natural language **E**xplanation model (MCLE) for VQA with semantic-level, image-level, and instance-level factual and counterfactual samples. MCLE extracts discriminative features and aligns the feature spaces from explanations with visual question and answer to generate more consistent explanations. We conduct extensive experiments, ablation analysis, and case study to demonstrate the effectiveness of our method on two VQA-NLE benchmarks.

Introduction

Deep neural networks have achieved significant progress on visual question answering (VQA) (Antol et al. 2015). However, most of them are black-box systems, which makes it hard to gain users' trust. It is a critical problem for these models to explain their decision-making process. In recent years, there has been some development of explainable VQA systems (Patro et al. 2019; Chen and Zhao 2022). Visualization analytic approaches generate a heatmap with different values by exploiting attention mechanisms and gradient analysis (Lu et al. 2016; Selvaraju et al. 2017) where the regions with higher values contribute the most to the predicted answers. However, such visualizations do not explain how these regions support the answer.

In contrast, natural language explanation (NLE) can provide the decision-making process of the model for users by generating a natural language sentence (Camburu et al. 2018; Park et al. 2018), which is more accessible to understand. NLE can also improve the ability of large models to perform complex reasoning (Wei et al. 2022). Post-hoc NLE methods have achieved good performance on VQA (Park et al. 2018; Kayser et al. 2021; Wu and Mooney 2019a). They first gain answers for VQA by exploiting a vision-language (VL) model. Then the predicted answers along with the visual questions are fed into an explanation generation model to gain corresponding explanations. To reduce the high storage and memory requirements caused by the addition of the task model in post-hoc NLE methods, a self-rationalization method is proposed to predict an answer and explain it by formulating the answer prediction as a text generation task along with the explanation (Sammani, Mukherjee, and Deligiannis 2022).

Despite their success, few VQA-NLE methods consider the faithfulness of the generated explanations and they suffer from the issue of logical inconsistency. As shown in Figure 1, we manually inspected a large number of VQA-NLE samples with wrong answer predictions generated by the NLX-GPT model (Sammani, Mukherjee, and Deligiannis 2022) and found that: (a) the relationships between generated explanations and answers are deductive unsatisfiability (the generated explanation does not logically lead to the answer); (b) the explanations are inconsistent with the facts in corresponding images (the model falsifies its counterfactual explanations for the answer without considering the visual information); and (c) the generated explanations are insensitivity with semantic perturbations on visual questions (the model fails to recognize the semantic change caused by changing only several words or visual objects). These findings raise fundamental questions on the role of explanations in VQA: *How to improve the faithfulness of explanations and reduce inconsistencies between explanations and visual question answers?*

To address the above challenges, in this paper, we propose a **M**ulti-level **C**ontrastive **L**earning based natural language **E**xplanation (MCLE) framework which can learn discriminative representations from semantic-level, image-level, and instance-level factual and counterfactual samples. MCLE can encourage faithful explanations to be close to their corresponding visual questions and answers while to be far from other counterfactual (negative) samples. Specif-

*Corresponding author

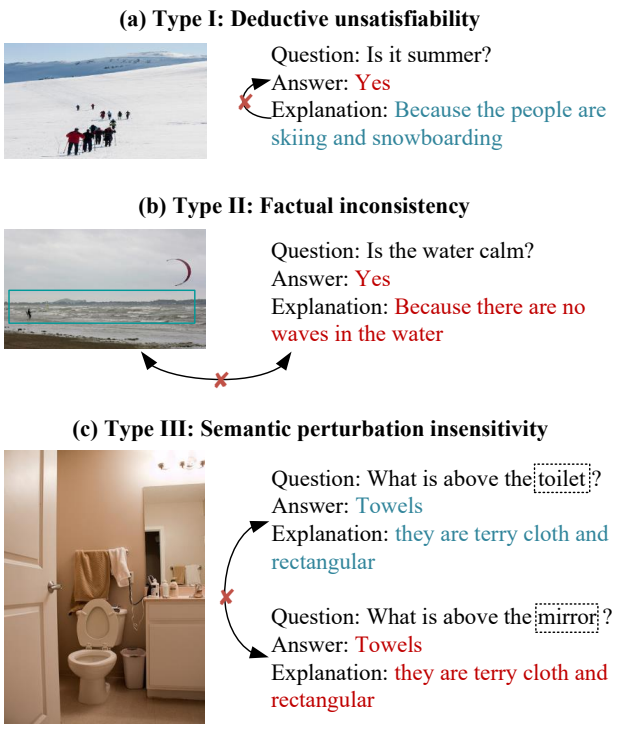

Figure 1: Three types of logical errors in VQA-NLE: (a) The generated explanation does not logically lead to the answer; (b) The model falsifies its counterfactual explanation for the answer without considering the facts in image; and (c) The model infers the same explanation and answer for a statement and its opposite semantics.

ically, MCLE consists of a vision-language (VL) model and a multi-level contrastive learning (CL) network. In our VL model, different from previous works (Sammani, Mukherjee, and Deligiannis 2022; Suo et al. 2023), we consider the VQA-NLE task as a chain-of-thought (COT) generation task (Wei et al. 2022), where the answer is produced after the explanation. In our multi-level CL network, three core modules are designed to learn high-quality representations to guide the model to generate faithful explanation, i.e., the semantic-level CL (SemanticCL) for deductive satisfiability, the image-level CL (ImageCL) for factual consistency, and the instance-level CL (InstanceCL) for semantic perturbation sensitivity. Powered with the COT strategy and the multi-level CL, the MCLE effectively models the logical relationships and promotes the logical consistency between explanations and visual question answers. In terms of both automatic measures and human evaluation, our MCLE outperforms the state-of-the-art models for the VQA-NLE task on two widely used datasets, and improves the faithfulness of the generated explanations.

In summary, we make the following contributions:

- We propose a multi-level contrastive learning (MCLE) framework, i.e., semantic-level, image-level, and instance-level, to perform discriminative representation learning, which improves logical consistency and faithfulness over VQA-NLE generation task.
- We propose a chain-of-thought generation strategy in the vision-language model for VQA-NLE, which boosts the accuracy of predicted answers while improves the reliability of generated explanations.
- The proposed MCLE achieves new state-of-the-art performance on VQA-X and A-OKVQA benchmark datasets. Ablation analysis and case study are conducted to help understand the working of MCLE.

Related Work

Explainability in VQA Given a question about an image, the goal of visual question answering is to generate an answer from both text and image information. It is firstly proposed by (Malinowski and Fritz 2014) and many approaches have been proposed such as joint embedding (Dong, Li, and Snoek 2018; Yao et al. 2019), attention mechanisms (Anderson et al. 2018; Lu et al. 2016), memory networks (Ma et al. 2018; Xiong, Merity, and Socher 2016) and graph neural networks (Kipf and Welling 2016; Velickovic et al. 2017). However, the reasoning process of the VQA models remains incomprehensible. Visualization technologies have been applied to achieve visual explanation (Selvaraju et al. 2017; Patro et al. 2019), but it has only limited expressiveness (Wu and Mooney 2019b). In contrast, text explanations formulated in (Park et al. 2018) are conducted on the VQA-NLE datasets and it utilizes human annotations to inspire the decision-making process of VQA models. (Kayser et al. 2021) combines a pre-trained language model and a VL model to generate free-text explanations while (Yang et al. 2022) uses stronger VL models (Li et al. 2020a) and generation models (Radford et al. 2019). (Sammani, Mukherjee, and Deligiannis 2022) proposes a unified model which can simultaneously predict answers and explanations based on a pre-trained caption model. Recently, (Suo et al. 2023) introduces a self-criticism strategy to model the logical relationship between answers and reasons. However, they still suffer from logical errors in VQA-NLE including deductive unsatisfiability, factual inconsistency, and semantic insensitivity.

Contrastive Learning The contrastive learning (CL) proposed in (Hadsell, Chopra, and LeCun 2006) has been extensively researched and shown impressive results in extracting powerful representations. Typical contrastive learning methods aim to learn representations by contrasting positive and negative pairs. Many researchers have attempted to incorporate the CL into their models in an unsupervised learning manner and they achieved great success. (Dosovitskiy et al. 2014) uses unlabeled instances for contrastive representation in visual recognition. (Khosla et al. 2020; Tian et al. 2020) utilize labeled data, benefitting tasks like VQA (Kim et al. 2021; Liang et al. 2020), image caption (Dai and Lin 2017; Li et al. 2020b), and visual grounding (Zhang et al. 2020). (Zhang, Zhang, and Xu 2021) employs multi-level CL for visual commonsense reasoning. Our approach adopts a multi-level contrastive architecture to improve the logical consistency and reliability over VQA-NLE explanation generation task.

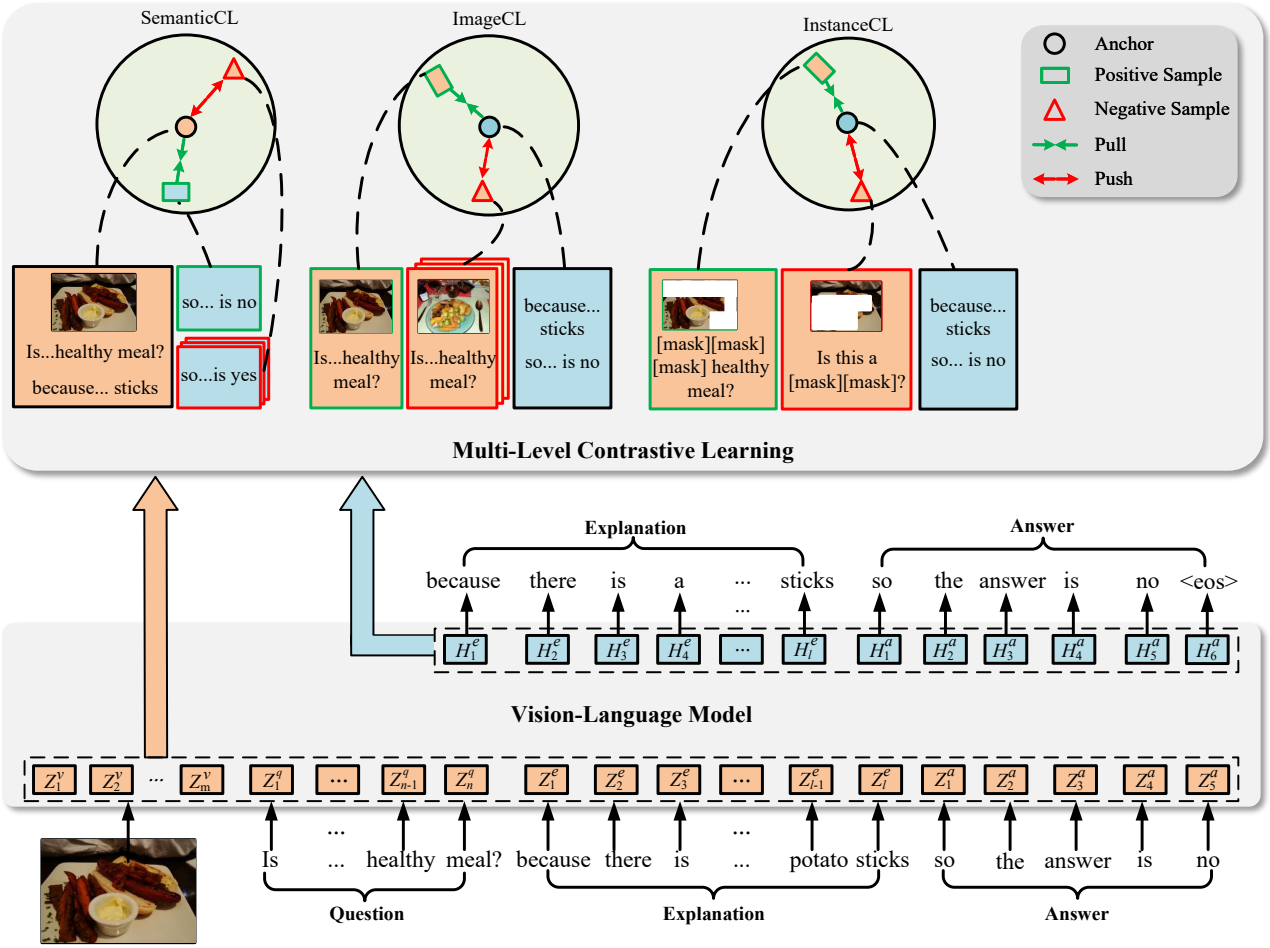

Figure 2: The overall architecture of MCLE. It consists of a vision-language model with chain-of-thought generation strategy and a multi-level contrastive learning network (semantic-level, image-level, and instance-level).

Method

In this section, we introduce our Multi-level Contrastive Learning based natural language Explanation (MCLE) framework. MCLE can improve the reliability of the rationales and strengthen the logical consistency between explanations and visual question answers. The overall architecture is shown in Figure 2 where the MCLE comprises a vision-language model and a multi-level contrastive learning network with semantic-level, image-level, and instance-level.

Vision-Language Model

Problem Formulation Given an image V and a natural language question $Q = (q_1, q_2, ..., q_n)$, where q_i represents the i-th word, the goal of VQA-NLE is to generate a corresponding free-text explanation with an answer.

Text and Image Representation Following previous works (Sammani, Mukherjee, and Deligiannis 2022), we adopt the GPT-2 (Radford et al. 2019) that pretrained on image caption task as our visual-language model and the CLIP (Radford et al. 2021) as our image encoder. The question features $Z_Q = (Z_1^q, Z_2^q, ..., Z_n^q)$ are obtained from the corresponding word embedding layer in GPT-2, where $Z_i^q \in \mathbb{R}^d$. The image features $Z_V = (Z_1^v, Z_2^v, ..., Z_m^v)$ are encoded by CLIP, where $Z_i^v \in \mathbb{R}^d$.

Chain-of-thought Generation To reduce the inconsistency between explanations and visual question answers, we introduce the chain-of-thought generation for VQA-NLE, which can mimic a rationale leading to the answer and provide an interpretable window into the decision-making progress. To inspire the model to generate faithful explanations and answers, the natural language '*because*' and '*so the answer is*' are as the prefixes of the ground-truth explanations and ground-truth answers, respectively. Then, like question features, the features of prefixed explanation and prefixed answer are obtained from the word embedding layer in GPT-2, where they are denoted by $Z_E = (Z_1^e, Z_2^e, ..., Z_l^e)$ and $Z_A = (Z_1^a, Z_2^a, ..., Z_5^a)$, respectively. By concatenating the prefixed explanation Z_E with prefixed answer Z_A, we get the chain-of-thought features $Z_T = [Z_E; Z_A]$. During training, all inputs (image Z_V, question Z_Q, chain-of-thought Z_T) are as a single sequence to the VL model. We train the VL model with the cross-

entropy objective to generate the chain-of-thought sequence $T = \{e_1, e_2, ..., e_l, a_1, a_2, ...a_5\}$ by minimizing:

$$\begin{aligned}\mathcal{L}_{vqa} &= -\log p\left(T \mid Z_V, Z_Q\right) \\ &= -\Big(\sum\nolimits_i \log p(e_i|H_i^e) + \sum\nolimits_i \log p(a_i|H_i^a)\Big)\end{aligned} \tag{1}$$

where

$$\begin{aligned}H_i^e &= \text{VL}(Z_V, Z_Q, Z_1^e, Z_2^e, ..., Z_{i-1}^e; \theta) \\ H_i^a &= \text{VL}(Z_V, Z_Q, Z_E, Z_1^a, Z_2^a, ..., Z_{i-1}^a; \theta)\end{aligned}$$

and θ denotes the parameters of the VL model. Note that, $H_1^e = \text{VL}(Z_V, Z_Q; \theta)$ and $H_1^a = \text{VL}(Z_V, Z_Q, Z_E; \theta)$.

Unlike other post-hoc methods, our explanation generation is solely based on visual questions and does not involve falsifying explanations based on the answer. Furthermore, the generated explanations can be used to prompt the generation of answers, which improves the logical consistency in VQA-NLE.

Multi-Level Contrastive Learning Network

Our multi-level contrastive learning network consists of three modules: SemanticCL, ImageCL, and InstanceCL. Following the contrastive learning framework in sequence to sequence learning (Lee, Lee, and Hwang 2020), we maximize the similarity between the pair of anchor and positive (factual) sequence, while minimize the similarity between the pair of anchor and negative (counterfactual) as follows:

$$\begin{aligned}\mathcal{L}_{CL} &= \text{CL}(\mathbf{x}, \mathbf{S}, \mathbf{y}) \\ &= -\log \frac{\exp\left(\text{sim}\left(\mathbf{e_x}, \mathbf{e_y}\right)/\tau\right)}{\sum_{\hat{\mathbf{x}} \in \mathbf{S}} \exp\left(\text{sim}\left(\mathbf{e}_{\hat{\mathbf{x}}}, \mathbf{e_y}\right)/\tau\right)}\end{aligned} \tag{2}$$

where

$$\begin{aligned}\mathbf{e_x} &= \xi\left(\mathbf{x}; \theta\right) \\ \xi\left([x_1, ..., x_t]; \theta\right) &= \text{AvgPool}([u_1, ..., u_t]) \\ u_t &= \text{ReLU}(\mathbf{W}x_t + \mathbf{b})\end{aligned}$$

and $\mathbf{x}$ denotes positive (factual) sample, $\mathbf{S}$ denotes the set of negative (counterfactual) samples, $\mathbf{y}$ denotes the anchor, and τ is the learned temperature parameter. The composition of affine transformation ξ with the ReLU and AvgPool projects the sequences $[x_1, ..., x_t] \in \mathbb{R}^{d \times t}$ onto the latent embedding space $\mathbf{e_x} \in \mathbb{R}^d$. The similarity $\text{sim}(\cdot, \cdot)$ measures between two sequences.

SemanticCL To guide our VQA-NLE model to generate explanations that logically lead to the answers, we design a semantic-level CL module (SemanticCL) to learn the relationship between explanations and answers. Specifically, the ground-truth answer sequence is taken as the positive sample, the random sampled K non-target answer sequences from the same batch are taken as the set of negative samples $\mathbf{S}$, and the combination of visual question and explanation is taken as the anchor. The contrastive loss in semanticCL is defined as follows:

$$\mathcal{L}_{semCL} = \text{CL}(\mathbf{x}, \mathbf{S}, \mathbf{y}) \tag{3}$$

where

$$\mathbf{x} = H_A,\ \mathbf{S} = \{\hat{H}_A\}_K,\ \mathbf{y} = [Z_V; Z_Q; Z_E]$$

and H_A, $\hat{H}_A$ denotes the positive and negative answer features obtained from the decoder hidden states of the VL model, respectively. Z_V, Z_Q, Z_E denotes the image, question, and explanation features, respectively, which are obtained from the image encoder and word embedding layer of the VL model. $\{\cdot\}_K$ denotes the set of K negative samples. Through the training, the corresponding explanation features are near to the ground-truth answer, while they are far away from the negative answers. In this way, the semanticCL helps learn the discriminative features between explanations and answers.

ImageCL The image-level CL module (ImageCL) aims to guide the model to generate explanations closely related to the visual information, rather than to falsify counterfactual explanations according to the question only. Specifically, the combination of explanation and answer $[H_E; H_A]$ is taken as the anchor $\mathbf{y}$. The original image with question $[Z_V; Z_Q]$ is taken as the factual sample $\mathbf{x}$. The counterfactual image with question $[\hat{Z}_V; Z_Q]$ is taken as the counterfactual sample.

During the counterfactual image sampling, we first calculate the score between original sample and other samples in the dataset by:

$$score = \text{sim}\left(\mathbf{e_{\hat{q}}}, \mathbf{e_q}\right) - \text{sim}\left(\mathbf{e_{\hat{a}}}, \mathbf{e_a}\right)$$

where $\mathbf{e_q}$ ($\mathbf{e_{\hat{q}}}$) and $\mathbf{e_a}$ ($\mathbf{e_{\hat{a}}}$) are the latent embeddings of question and answer in original (counterfactual) sample, respectively.

Then we select the images of top-K samples with the highest scores as the counterfactual images, which have a similar question but a different answer from the original sample, to guide the model to perceive the visual contents and eliminate the factual inconsistency caused by language bias.

The contrastive loss in our ImageCL is defined as follows:

$$\mathcal{L}_{imgCL} = \text{CL}(\mathbf{x}, \mathbf{S}, \mathbf{y}) \tag{4}$$

where

$$\mathbf{x} = [Z_V; Z_Q],\ \mathbf{S} = \{[\hat{Z}_V; Z_Q]\}_K,\ \mathbf{y} = [H_E; H_A]$$

Through the training, the explanations are near to the corresponding image, while they are far away from the counterfactual images. In this way, the ImageCL helps learn the discriminative features between explanations and images.

InstanceCL To help VQA-NLE model perceive the semantic changes caused by fine-grained visual or text perturbations, we design an instance-level CL module (InstanceCL). In this module, a gradient-based counterfactual transformation strategy is adopted to synthesize factual and counterfactual samples. We apply the modified Grad-CAM (Selvaraju et al. 2017) to derive the contribution of the i-th object and the j-th word to answer by the following functions:

$$s(a, Z_i^v) = S(P_{vqa}(a), Z_i^v) = (\nabla_{Z_i^v} P_{vqa}(a))^\top \mathbf{1} \tag{5}$$

$$s(a, Z_j^q) = S(P_{vqa}(a), Z_j^q) = (\nabla_{Z_j^q} P_{vqa}(a))^\top \mathbf{1} \tag{6}$$

where $P_{vqa}(a)$ is the predicted answer probability of the ground truth answer, Z_i^v is the i-th object features, Z_j^q is the j-th word features, and $\mathbf{1}$ is an all ones vector.

Obviously, if the score $s(a, \cdot)$ is higher, the contribution of the object Z_i^v (or Z_j^q) to the answer is larger. Based on such contribution scores, the top-K objects and words with the highest scores are collected as the factual samples $[Z_V^+; Z_Q^+]$ while the counterfactual samples $[Z_V^-; Z_Q^-]$ are generated by masking the corresponding factual samples. The contrastive loss in InstanceCL is defined as follows:

$$\mathcal{L}_{insCL} = \mathrm{CL}(\mathbf{x}, \mathbf{S}, \mathbf{y}) \tag{7}$$

where

$$\mathbf{x} = [Z_V^+; Z_Q^+],\ \mathbf{S} = [Z_V^-; Z_Q^-],\ \mathbf{y} = [H_E; H_A]$$

and the union set $Z_V = Z_V^+ \cup Z_V^-$, and $Z_Q = Z_Q^+ \cup Z_Q^-$. During training, explanations are near to the corresponding objects and words, while they are far away from unrelated objects and words. In this way, the ImageCL helps perceive the key fine-grained content in the image and question.

Overall Loss

The overall loss of our MCLE is:

$$\mathcal{L} = \mathcal{L}_{vqa} + \alpha\mathcal{L}_{semCL} + \beta\mathcal{L}_{imgCL} + \gamma\mathcal{L}_{insCL} \tag{8}$$

where α, β, and γ are the trade-off parameters. MCLE improves the logical consistency and reliability of the VQA-NLE task by jointly optimizing the main loss (the vision-language model) and three auxiliary losses (the multi-level contrastive learning network).

Experiment

Datasets

Following (Suo et al. 2023), we conduct empirical experiments on two widely used VQA-NLE benchmarks.

VQA-X (Park et al. 2018) is collected from the VQA dataset (Antol et al. 2015) and provides additional explanations for the answers. It consists of 28K images and 33K QA pairs, split into 29K/1.4K/1.9K for training, validation, and testing, respectively.

A-OKVQA (Schwenk et al. 2022) is collected from the COCO dataset (Lin et al. 2014). It includes about 25K Question/Answer/Rationale triplets, split into 17.1K/1.1K/6.7K for training, validation, and testing, respectively.

Evaluation Measures

Automatic Evaluation Following (Suo et al. 2023), the generated explanations are evaluated in terms of the metrics BLUE (Papineni et al. 2002), METEOR (Denkowski and Lavie 2014), ROUGE-L (Lin 2004), SPICE (Anderson et al. 2016), and CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015). The predicted answers are evaluated in terms of the metric Accuracy.

Human Evaluation Following (Suo et al. 2023), we use human evaluations to measure the faithfulness and logicality of the explanations, since they are not always reflected by the automatic metrics (Kayser et al. 2021). Specifically, three human evaluators are employed to determine each generated explanation whether it is consistent with the answer and they select an option from [yes, weak yes, weak no, no], corresponding to scores [1, 2/3, 1/3, 0], respectively. We compute an average among total scores of all test samples to obtain the final human evaluation score. Meanwhile, these evaluators are asked to choose the reason for unqualified explanations: deductive unsatisfiability, factual inconsistency, and semantic perturbation insensitivity (see Figure 1).

Experimental Setup

Baselines We compare with five strong baselines.

- **PJ-X** (Park et al. 2018) is a post-hoc method by creating attention features to guide the generation of textual explanations.
- **FME** (Wu and Mooney 2019a) uses the Grad-CAM (Selvaraju et al. 2017) to generate the explanations which can be traced back to the relevant object set.
- **e-UG** (Kayser et al. 2021) generates the explanations by combining UNITER (Chen et al. 2020) and GPT-2.
- **NLX-GPT** (Sammani, Mukherjee, and Deligiannis 2022) can simultaneously predict an answer and explain it by formulating the answer prediction as a text generation task along with the explanation.
- **S^{3}C** (Suo et al. 2023) is a self-critical VQA-NLE method that can model the logical relationships between answer-explanation pairs.

Implementation We conduct all experiments on NVIDIA GTX 3080 Ti GPUs with PyTorch 1.9.0. We take the GPT-2 model that pre-trained on image caption task (Sammani, Mukherjee, and Deligiannis 2022) as our vision-language model backbone. The temperature τ is set to 0.2. The hyperparameters top-K in the multi-level CL (SemanticCL, ImageCL, and InstanceCL) are set to (3, 3, 2), respectively. The trade-off parameters α, β, and γ are set to 0.1, 0.2, and 0.2, respectively. The maximum length of text sentence cuts at 40, batch size is 16, and training epoch is 30. See our released code at https://github.com/laichengen/MCLE.

Main Results on Automatic Evaluation

Unfiltered Scenario The performance comparison of different methods is shown in Table 1, where the best results are in boldface. We have the following observations. We observe that our MCLE achieves the new state-of-the-art performance on two VQA-NLE datasets (VQA-X and A-OKVQA). Specifically, our MCLE outperforms the best baseline with 2.4% improvement in terms of SPICE on the VQA-X dataset, while with 2.2% improvement in terms of CIDEr on the A-OKVQA. Furthermore, our MCLE gets the best result over all of the 12 automatic evaluation settings on the two datasets, with average 1.7% and 1.4% improvements respectively. This shows that our model can generate more reliable explanations. As for the accuracy of answers (see the column of "Acc"), our MCLE can simultaneously boost the precision of answers and corresponding explanations. These improvements of our MCLE over baselines could be

	VQA-X							A-OKVQA						
Approach	B4	M	R	C	S	Acc	Human	B4	M	R	C	S	Acc	Human
PJ-X	19.5	18.2	43.4	71.3	15.1	76.4	65.4	-	-	-	-	-	-	-
FME	24.4	19.5	47.4	88.8	17.9	75.5	-	-	-	-	-	-	-	-
e-UC	-	-	-	-	-	-	-	15.1	18.1	42.4	51.5	14.9	25.6	44.1
NLX-GPT	25.6	21.5	48.7	97.2	20.2	83.1	70.2	20.1	17.0	46.3	65.4	15.8	28.7	46.9
S^3C	27.8	22.8	50.7	104.4	21.5	85.6	77.4	22.5	18.5	48.4	74.4	18.1	33.5	54.7
MCLE (ours)	**28.6**	**24.2**	**52.3**	**106.7**	**23.9**	**87.7**	**80.8**	**23.1**	**19.7**	**50.1**	**76.6**	**19.7**	**34.7**	**57.9**

Table 1: Comparison with the state-of-the-art methods on the VQA-X and A-OKVQA datasets in the scenario of "unfiltered" scores. ("unfiltered" indicates that the explanations are evaluated regardless of whether the answer is true or false, while "filtered" is to only consider the explanations that have correct answers.) The B4, M, R, C, S, Acc, and Human are short for BLEU-4, METEOR, ROUGE-L, CIDEr, SPICE, Answer Accuracy, and Human Evaluation, respectively.

	VQA-X						
Approach	B4	M	R	C	S	Acc	Human
PJ-X	22.7	19.7	46.0	82.7	17.1	76.4	69.3
FME	23.1	20.4	47.1	87.0	18.4	75.5	-
e-UG	23.2	22.1	45.7	74.1	20.1	80.5	71.4
NLX-GPT	28.5	23.1	51.5	110.6	22.1	83.1	73.7
S^3C	30.7	23.9	52.1	116.7	23.0	85.6	79.2
MCLE (ours)	**31.2**	**24.2**	**53.1**	**118.3**	**24.2**	**87.7**	**81.3**

Table 2: Comparison with the state-of-the-art methods on the VQA-X dataset in the scenario of "filtered" scores. ("unfiltered" indicates that the explanations are evaluated regardless of whether the answer is true or false, while "filtered" is to only consider the explanations that have correct answers.)

	VQA-X		
Approach	Type I	Type II	Type III
PJ-X	28.4%	21.1%	9.2%
e-UG	25.4%	22.8%	8.7%
NLX-GPT	22.2%	20.3%	9.1%
S^3C	18.9%	17.3%	8.2%
MCLE (ours)	**17.1%**	**15.7%**	**7.8%**

Table 3: The main reason of unqualified explanations on the VQA-X dataset. Three types of logical errors: (a) Type I: Deductive unsatisfiability, (b) Type II: Factual inconsistency, and (c) Type III: Semantic perturbation insensitivity (see Figure 1).

attributed to two reasons: i) MCLE adopts an explain-then-predict framework with chain-of-thought generation strategy, which can mimic a rationale leading to the answer; and ii) The multi-level contrastive learning network is able to learn high-quality representations to guide the model to generate logical consistency explanations.

Filtered Scenario To verify the algorithm's validity, we follow (Kayser et al. 2021) to report the filtered scores on the VQA-X dataset as shown in Table 2. Our MCLE achieves a new state-of-the-art on VQA-X. Specifically, our method can outperform the baseline methods with 1.6% improvement in terms of SPICE and with 1.4% improvement in terms of answer accuracy.

Main Results on Human Evaluation

Unfiltered and Filtered Scenarios We conduct the human evaluation to evaluate the correctness and faithfulness of generated explanations. As shown in Table 1, compared to other five methods, the Human score of our MCLE is improved by 3.4 points on VQA-X in the scenario of unfiltered scores. As shown in Table 2, compared to other five methods, the Human score of our MCLE is improved by 2.1 points on VQA-X in the scenario of filtered scores.

Logical Errors Moreover, we also ask the human evaluators to select the reasons for each unqualified explanation on the VQA-X dataset. As shown in Table 3, the insufficient explanations caused by deductive unsatisfiability are reduced by 1.8%, the irrelevant explanations caused by factual inconsistency are reduced by 1.6%, and the meaningless explanations caused by semantic perturbation insensitivity are reduced by 0.4%. These results indicate that our MCLE can obtain relatively better rationales and empirically confirm the effectiveness of our method.

Ablation Studies

The ablation results of the full MCLE model and its five variants are shown in Table 4. From the results, we have the following findings.

Firstly, for the effectiveness of the chain-of-thought (COT) generation strategy which mimics a rationale leading to the answer and provides an interpretable window into the decision-making progress of the model, MCLE w/o CTG performs worse than MCLE. For example, CIDEr reduces by 1.2% on VQA-X. It verifies that the COT strategy is important for the model to improve the explanation's logical consistency.

Secondly, for the effectiveness of the SemanticCL strategy which learns the discriminative features between explanations and answers, MCLE w/o SemanticCL performs worse than MCLE. For example, SPICE reduces by 1.6% on VQA-X. It verifies that the SemanticCL is important to guide the model to generate explanations that logically lead to the answers.

Thirdly, for the effectiveness of the ImageCL strategy which learns the discriminative features between explanations and images, MCLE w/o ImageCL performs worse than

	VQA-X							A-OKVQA						
Approach	B4	M	R	C	S	Acc	Human	B4	M	R	C	S	Acc	Human
MCLE	**28.6**	**24.2**	**52.3**	**106.7**	**23.9**	**87.7**	**80.8**	**23.1**	**19.7**	**50.1**	**76.6**	**19.7**	**34.7**	**57.9**
w/o COT	27.8	23.6	51.7	105.5	23.1	86.5	79.4	22.5	18.8	48.6	75.7	19.1	33.7	56.2
w/o SemanticCL	27.1	23.5	51.1	104.6	22.3	86.1	78.8	22.3	18.2	48.6	75.9	18.8	33.1	55.8
w/o ImageCL	26.9	21.5	48.7	103.3	21.7	85.8	77.6	20.3	17.8	47.8	75.4	16.8	31.7	53.9
w/o InstanceCL	27.3	22.8	49.9	105.1	22.8	86.4	79.2	21.8	18.1	49.4	75.5	18.3	32.9	54.6
w/o All	25.9	22.1	48.3	98.1	20.9	83.5	71.8	20.1	16.8	46.2	66.5	16.2	28.7	47.2

Table 4: Ablated results of our MCLE and its key components, chain-of-thought (COT) generation strategy and the multi-level contrastive learning network (semantic-level, image-level, and instance-level).

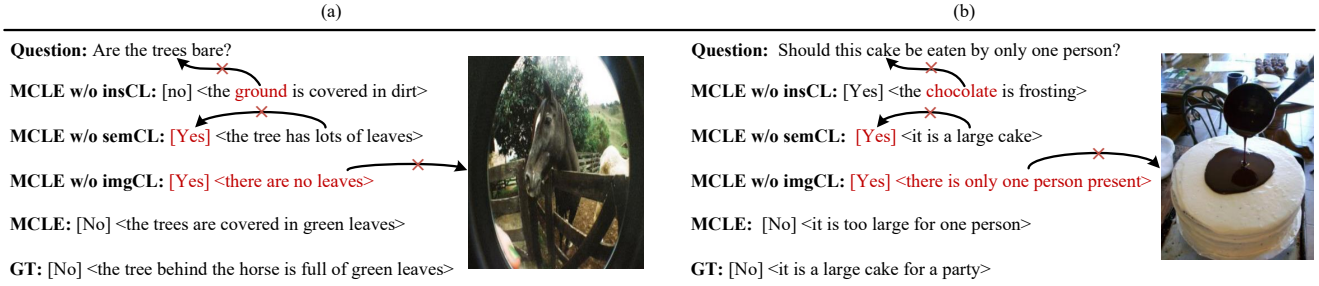

Figure 3: Case study on the generated explanations on the VQA-X dataset. The $[\cdot]$ and $< \cdot >$ indicate answers and explanations respectively. We show the results of our full MCLE model and its three variants. GT denotes by the ground truth.

MCLE. For example, ROUGE-L reduces by 3.6% on VQA-X. It verifies that the ImageCL is important to guide the model to generate explanations closely related to the visual information, rather than falsifying counterfactual explanations caused by language bias.

Fourthly, for the effectiveness of the InstanceCL strategy which perceives the semantic changes caused by fine-grained visual and text perturbation, MCLE w/o InstanceCL performs worse than MCLE. For example, ROUGE-L reduces by 2.4% on VQA-X. It verifies that the InstanceCL is important to guide the model to perceive the key fine-grained content in image and question.

Obviously, MCLE w/o All is the worst. For example, CIDEr reduces by 8.6% on VQA-X. It further shows that both of the chain-of-thought (COT) generation strategy and multi-level contrastive learning network in our MCLE contribute to the performance improvements.

Case Studies

We have quantitatively demonstrated the effectiveness of our MCLE by comparing with five state-of-the-art methods, and conducted detailed ablation study on the contributions from the core components of chain-of-thought generation strategy and multi-level contrastive learning network. In this section, to obtain an intuitive understanding of how the proposed MCLE works, we show typically qualitative results from the NLX-GPT module (Sammani, Mukherjee, and Deligiannis 2022) on the VQA-X dataset by comparing the generated explanations and answers of MCLE and its variants.

As shown in Figure 3(a), the MCLE w/o InstanceCL generates an explanation that is relevant to the word "ground" rather than "trees", maybe caused by the language bias in the dataset (e.g., "Are the ground bare?"). For the MCLE w/o SemanticCL, although the generated explanation correctly identifies the tree having lots of leaves, the predicted answer is wrongly "yes" (answering the trees as bare for the question), which is contradictory to the explanation and suffers from deductive unsatisfiability (Type I logical error in Figure 1). The MCLE w/o ImageCL falsifies the explanation that the tree in the image has no leaves, which is inconsistent with the factual image and suffers from factual inconsistency (Type II logical error). For our full MCLE model with all three-level contrastive learning (CL) components, it correctly generates more faithful rationales and predicts logically consistent answers. This shows that the multi-level CL can help vision-language models improve the logical consistency between explanations and visual question answers. Figure 3(b) is another similar case to help understand the working of our proposed MCLE framework and the contributions from the three-level CL network.

Conclusion

We proposed a novel self-supervised Multi-level Contrastive Learning based natural language Explanation model (MCLE) for VQA with semantic-level, image-level, and instance-level factual and counterfactual (negative) samples. MCLE can learn discriminative features and align the feature spaces from explanations with visual questions and answers to generate more consistent explanations. Our model improves consistent and faithful explanations while reduces the deductive unsatisfiability, factual inconsistency, and semantic perturbation insensitivity. From both automatic measures and human evaluations, our MCLE achieves a new state-of-the-art on VQA-NLE task.

Acknowledgments

This work is supported by National Natural Science Foundation of China (No. 62306220).

References

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. SPICE: semantic propositional image caption evaluation. In *European Conference on Computer Vision*, 382–398.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. 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*, 6077–6086.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2425–2433.

Camburu, O.-M.; Rocktäschel, T.; Lukasiewicz, T.; and Blunsom, P. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.

Chen, S.; and Zhao, Q. 2022. Rex: Reasoning-aware and grounded explanation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15586–15595.

Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, 104–120.

Dai, B.; and Lin, D. 2017. Contrastive learning for image captioning. *Advances in Neural Information Processing Systems*, 30.

Denkowski, M.; and Lavie, A. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, 376–380.

Dong, J.; Li, X.; and Snoek, C. G. 2018. Predicting visual features from text for image and video caption retrieval. *IEEE Transactions on Multimedia*, 20(12): 3377–3388.

Dosovitskiy, A.; Springenberg, J. T.; Riedmiller, M.; and Brox, T. 2014. Discriminative unsupervised feature learning with convolutional neural networks. *Advances in neural information processing systems*, 27.

Hadsell, R.; Chopra, S.; and LeCun, Y. 2006. Dimensionality Reduction by Learning an Invariant Mapping. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Volume 2*, 1735–1742.

Kayser, M.; Camburu, O.-M.; Salewski, L.; Emde, C.; Do, V.; Akata, Z.; and Lukasiewicz, T. 2021. 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*, 1244–1254.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33: 18661–18673.

Kim, S.; Jeong, S.; Kim, E.; Kang, I.; and Kwak, N. 2021. Self-supervised pre-training and contrastive representation learning for multiple-choice video qa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 13171–13179.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Lee, S.; Lee, D. B.; and Hwang, S. J. 2020. Contrastive Learning with Adversarial Perturbations for Conditional Text Generation. In *International Conference on Learning Representations*.

Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. 2020a. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In *European Conference on Computer Vision*, 121–137.

Li, Z.; Tran, Q.; Mai, L.; Lin, Z.; and Yuille, A. L. 2020b. Context-aware group captioning via self-attention and contrastive features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3440–3450.

Liang, Z.; Jiang, W.; Hu, H.; and Zhu, J. 2020. Learning to contrast the counterfactual samples for robust visual question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, 3285–3292.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision*, 740–755.

Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2016. Hierarchical question-image co-attention for visual question answering. *Advances in neural information processing systems*, 29.

Ma, C.; Shen, C.; Dick, A.; Wu, Q.; Wang, P.; Van den Hengel, A.; and Reid, I. 2018. Visual question answering with memory-augmented networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6975–6984.

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

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Park, D. H.; Hendricks, L. A.; Akata, Z.; Rohrbach, A.; Schiele, B.; Darrell, T.; and Rohrbach, M. 2018. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8779–8788.

Patro, B. N.; Lunayach, M.; Patel, S.; and Namboodiri, V. P. 2019. U-cam: Visual explanation using uncertainty based class activation maps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7444–7453.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Sammani, F.; Mukherjee, T.; and Deligiannis, N. 2022. 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*, 8322–8332.

Schwenk, D.; Khandelwal, A.; Clark, C.; Marino, K.; and Mottaghi, R. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, 146–162.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Suo, W.; Sun, M.; Liu, W.; Gao, Y.; Wang, P.; Zhang, Y.; and Wu, Q. 2023. S3C: Semi-Supervised VQA Natural Language Explanation via Self-Critical Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2646–2656.

Tian, Y.; Sun, C.; Poole, B.; Krishnan, D.; Schmid, C.; and Isola, P. 2020. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33: 6827–6839.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.

Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y.; et al. 2017. Graph attention networks. *stat*, 1050(20): 10–48550.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837.

Wu, J.; and Mooney, R. 2019a. Faithful Multimodal Explanation for Visual Question Answering. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 103–112.

Wu, J.; and Mooney, R. 2019b. Self-critical reasoning for robust visual question answering. *Advances in Neural Information Processing Systems*, 32.

Xiong, C.; Merity, S.; and Socher, R. 2016. Dynamic memory networks for visual and textual question answering. In *International conference on machine learning*, 2397–2406.

Yang, Q.; Li, Y.; Hu, B.; Ma, L.; Ding, Y.; and Zhang, M. 2022. Chunk-aware alignment and lexical constraint for visual entailment with natural language explanations. In *Proceedings of the 30th ACM International Conference on Multimedia*, 3587–3597.

Yao, T.; Pan, Y.; Li, Y.; and Mei, T. 2019. Hierarchy parsing for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2621–2629.

Zhang, X.; Zhang, F.; and Xu, C. 2021. Multi-level counterfactual contrast for visual commonsense reasoning. In *Proceedings of the 29th ACM International Conference on Multimedia*, 1793–1802.

Zhang, Z.; Zhao, Z.; Lin, Z.; He, X.; et al. 2020. Counterfactual contrastive learning for weakly-supervised vision-language grounding. *Advances in Neural Information Processing Systems*, 33: 18123–18134.