Variational Causal Inference Network for Explanatory Visual Question Answering

Dizhan Xue\textsuperscript{1,2} Shengsheng Qian\textsuperscript{1,2} Changsheng Xu\textsuperscript{1,2,3}
\textsuperscript{1}State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences
\textsuperscript{2}University of Chinese Academy of Sciences
\textsuperscript{3}Peng Cheng Laboratory
xuedizhan17@mails.ucas.ac.cn, \{shengsheng.qian, csxu\}@nlpr.ia.ac.cn

Abstract

Explanatory Visual Question Answering (EVQA) is a recently proposed multimodal reasoning task that requires answering visual questions and generating multimodal explanations for the reasoning processes. Unlike traditional Visual Question Answering (VQA) which focuses solely on answering, EVQA aims to provide user-friendly explanations to enhance the explainability and credibility of reasoning models. However, existing EVQA methods typically predict the answer and explanation separately, which ignores the causal correlation between them. Moreover, they neglect the complex relationships among question words, visual regions, and explanation tokens. To address these issues, we propose a Variational Causal Inference Network (VCIN) that establishes the causal correlation between predicted answers and explanations, and captures cross-modal relationships to generate rational explanations. First, we utilize a vision-and-language pretrained model to extract visual features and question features. Secondly, we propose a multimodal explanation gating transformer that constructs cross-modal relationships and generates rational explanations. Finally, we propose a variational causal inference to establish the target causal structure and predict the answers. Comprehensive experiments demonstrate the superiority of VCIN over state-of-the-art EVQA methods.

1. Introduction

Multimodal reasoning is a vital ability for humans and a fundamental problem for artificial intelligence [27, 39, 8]. Despite the promising performance of deep neural networks on various multimodal reasoning tasks [35, 37, 47, 34, 36], existing models typically generate reasoning results without explaining the rationale behind their results. Consequently, the low explainability of the generated results severely reduces the credibility and restricts the application of reasoning models. To address this issue, Chen and Zhao [11] recently proposed Explanatory Visual Question Answering (EVQA) task, which expands upon Visual Question Answering (VQA) [5, 15] by requiring multimodal reasoning explanations. As shown in Figure 1, while traditional VQA aims to answer a question with a related image, EVQA additionally requires explaining the reasoning process.

Figure 1. An example of Visual Question Answering (VQA) and Explanatory Visual Question Answering (EVQA): VQA requires answering the question with a related image, while EVQA additionally requires explaining the reasoning process.
of-the-art EVQA method REX [11] predicts the answer “Microwave” but explains the target object as “phone” in this example. These contradictory results cannot be adopted in practice and even hurt the credibility of the reasoning system. Besides the qualitative analysis, we compute the Consistency score (see Section 5.2) of the results inferred by REX on GQA-REX dataset, which is 74.69% and far from 100% of the ideal results. These analyses jointly reveal the low answer-explanation consistency of the existing methods. Therefore, we have to address Challenge 1: How to establish the consistency relation between the predicted answers and explanations to improve the credibility of the EVQA model?

Moreover, the existing methods [11] for EVQA feed a fused input feature into LSTM-based decoders to generate the explanation while ignoring the complex relationships among question words, visual regions, and explanation tokens. However, an ideal explanation is usually similar to the visual question in terms of both semantics and sentence structure. For the example in Figure 1, the explanation simply transforms the question (an interrogative sentence) into a declarative sentence while replacing nouns with specific visual regions. Furthermore, since the consistency between explanations and answers is essential, the quality of the generated explanations can also affect the accuracy of the predicted answers. Therefore, we need to address Challenge 2: How to construct relationships among question words, visual regions, and explanation tokens to improve the quality of the generated multimodal explanation for EVQA?

Motivated by the above observations, we propose a novel Variational Causal Inference Network (VCIN) for EVQA to improve both the quality and consistency of the inferred results. For Challenge 1, we propose a variational causal inference to establish the causal correlation between answer and explanation while reasoning, which can significantly improve the consistency between the predicted answers and explanations. Different from the existing work of causal learning in CV and NLP [24, 23, 48] that typically focuses on eliminating biased dependency, we propose to establish the ignored causal correlation in the Structural Causal Model (SCM) for EVQA. Additionally, we propose an automatic Consistency (Con.) metric to evaluate the answer-explanation consistency and facilitate the research of credible reasoning for EVQA. For Challenge 2, we design a multimodal explanation gating transformer to capture complex relationships among question words, visual regions, and explanation tokens, which can generate coherent and rational explanations of the reasoning processes. To flexibly generate multimodal explanations, we adopt a multimodal gating network to dynamically select word tokens and visual tokens for explanation generation. Comprehensive experiments on EVQA benchmark datasets demonstrate a significant performance boost of our proposed model compared with the state-of-the-art methods. In brief, the contributions of this paper are listed as follows:

- We propose an end-to-end Variational Causal Inference Network (VCIN) by converting the target SCM into deep variational inference and designing a multimodal explanation gating transformer to improve both the credibility and quality of the inferred results for Explanatory Visual Question Answering (EVQA).
- We propose a novel variational causal inference to establish the causal correlation between answer and explanation while reasoning, which can significantly improve the answer-explanation consistency of the predicted results. Additionally, we propose an automatic metric named Consistency (Con.) to evaluate the answer-explanation consistency and facilitate the research of credible reasoning for EVQA.
- We design a multimodal explanation gating transformer to capture complex relationships among question words, visual regions, and explanation tokens for generating rational multimodal explanations of reasoning processes. Additionally, we utilize a multimodal gating network to flexibly generate visual and textual tokens in explanations.
- Extensive experiments on EVQA benchmark datasets indicate the superiority of the proposed method compared with the state-of-the-art methods in terms of both the quality and consistency of the inferred results.

2. Related Work

2.1. Explanatory Visual Question Answering

Explanatory Visual Question Answering (EVQA) [11] is a recently proposed task. While Visual Question Answering (VQA) requires answering a question with a related image [5, 15, 41], EVQA additionally aims at generating user-
friendly explanations of the reasoning process to improve the explainability of the inferred results.

Existing VQA methods mainly focus on effectively learning features of images and questions, and fusing multimodal features for answer prediction. For image representation, grid features [16, 10, 9] extracted by ResNet [13], ResNetXt [46], or ViT [12] and object features [4, 50, 28] extracted by Faster R-CNN [40] are two widely-used options. For question representation, Glove [32] and BERT [31] are two widely-used options. Moreover, we propose a variational causal inference to improve the consistency between the predicted answers and explanations. For question representation, Glove [32] and BERT [31] are two widely-used options. Moreover, we propose a variational causal inference to improve the consistency between the predicted answers and explanations.

Figure 3. Structural causal models for Explanatory Visual Question Answering. Q denotes the textual question, I denotes the question-related image, M denotes multimodal content features, E and E' denote the explanation, F denotes robust explanation feature, and A denotes the answer. Gray nodes correspond to observed/input features, white nodes correspond to inferred features.

2.2. Causal Inference

Recently, researchers have incorporated causal inference into deep learning models [26, 52, 24]. These efforts enable DNNs to learn causal effects, which improves the performance of models for various applications, such as semantic segmentation [52], image caption [24], and sequential prediction [48]. For example, Zhang et al. [52] propose a context adjustment in the structural causal model to remove the confounding bias in image-level classification. Causal inference has also been introduced to VQA. For instance, Agarwal et al. [2] propose automated semantic image manipulations to alleviate spurious correlations while learning. Niu et al. [29] propose a counterfactual inference framework to capture and mitigate language bias in VQA. Yang et al. [51] propose a causal attention mechanism to reduce the confounding bias which can mislead attention modules.

Different from the existing work that typically focuses on eliminating biased dependency in learning, we propose to reconstruct the ignored causal correlation between explanation and answer for EVQA by converting the target structural causal model into a deep variational inference.

3. Notation and Problem Formulation

We first introduce some notations used in this paper. EVQA aims to predict the answer of a given question with a related image and generate a multimodal explanation of the reasoning process. The input can be denoted as (Q, I). Q = q1q2...qm is a textual question of total m words and qi is the ith word in the question. I is a question-related image that can be represented by an RGB tensor. The output of an EVQA model can be denoted as (A, E). A is the answer selected from a predefined set {c1, ..., cK} of total K answers. E = c1c2...cn is the explanation of total n tokens, where ci is the ith token that can be a word from a predefined vocabulary or a number linking to a visual region. In training, ground truth outputs (A', E') are given.

The core of EVQA is to learn a multimodal reasoning model g : (Q, I) → (A, E), which can simultaneously answer the visual question and generate the explanation to conduct explainable and credible multimodal reasoning.

4. Methodology

Next, we introduce our causal inference method for EVQA. Due to the space limitation, some computation details are included in Supplementary Material.

4.1. Causal Perspective of EVQA

We first analyze the limitations of traditional EVQA models and introduce our solution from a causal perspective. We demonstrate the structural causal models (SCMs) [31] of different methods in Figure 3.

Traditional model As shown in Figure 3 (a), traditional EVQA methods [11] learn multimodal content features M based on input question Q and image I, which are further utilized to predict the explanation E and the answer A separately. While merely optimizing the marginal probabilities...
P(E = E'|M) and P(A = A'|M), the joint probability P(E = E', A = A'|M) that implies the consistency between E and A is ignored. Therefore, these methods usually lead to inconsistent explanations and answers.

**Ideal model** In ideal, as shown in Figure 3 (b), to maximize the accuracy of the predicted answer that is also consistent with the ground truth explanation, we can optimize the joint probability \( P(E', A = A'|M) = P(A = A'|M, E') P(E|M) \), where \( E' \) is the ground truth explanation and is observed in the ideal model, i.e., \( P(E'|M) = 1 \) is a Dirac delta distribution.

**Joint model** However, since \( E' \) is unavailable in the test, we propose an approximate model in Figure 3 (c), where the joint probability \( P(E = E', A = A'|M) = P(E = E'|M) P(A = A'|M, E = E') \) is optimized. However, since \( E \) is a token sequence, it is empirically hard to generate the identical ground truth in the test, i.e., \( E(M) = E' \). Therefore, optimizing \( P(A = A'|M, E = E') \) can hurt effectiveness and robustness in test where we can only utilize the explanation \( E = E' \) with the highest generative probability, i.e., \( E^* = \arg \max_E P(E|M) \).

**Our model** To alleviate the impact of the distribution shift between \( E' \) in training and \( E^* \) in test, as shown in Figure 3 (d), we add a front-door path \( E \rightarrow F \rightarrow A \) in the SCM, where \( F \) is a robust explanation feature. To improve the robustness of \( F \), we model \( F \) to follow a Gaussian distribution, i.e., \( F \sim N(\mu_E, diag(\sigma_E^2)) \), where \( \mu_E, \sigma_E^2 \) are two \( d_f \)-dimensional vectors computed based on \( E \). Moreover, we aim at minimizing the Kullback-Leibler (KL) divergence \( KL(P(F|E')||P(F|E^*)) \) to reduce the bias between distributions \( P(A|M, E') \) and \( P(A|M, E^*) \).

### 4.2. Variational Causal Inference

In this section, we introduce the optimization objectives of our method. We denote the distributions of our model and the ideal model as \( p \) and \( q \), respectively. Similar to Figure 3 (d), we add the explanation feature \( F \) to the ideal model. To train our model, our first objective is maximizing the likelihood of predicting the true answer \( A' \) while modeling causal effects from explanation to answer as follows:

\[
\log p(A'|M) \geq E_{p(F|M)} \log p(A'|M, F) + \log p(F|M) - \log q(F|M) \\
= p(F|M) \log p(A'|M, F) - KL(q(F|M) \parallel p(F|M)) \\
= E_{q(F|E')} \log p(A'|M, F) - KL(q(F|E') \parallel p(F|M)) ,
\]

where we utilize the following lemma:

\[
q(F|M) = \sum_E q(F|E)q(E|M) = q(F|E') ,
\]

since \( q(E|M) \) is a Dirac delta distribution satisfying \( q(E'|M) = 1. \) However, in Equation 1, \( p(F|M) = \sum_E p(F|E)p(E|M) \) is difficult to compute since computing \( \{p(E|M)\}_E \) is of exponential complexity and there is no explicit algorithm to sample \( E \sim p(E|M) \). Therefore, we propose to utilize an approximation \( P(F|E^*) \) which corresponds to the test scenario where \( E^* \) is the generated explanation. To sum up, we can obtain our variational causal inference loss as follows:

\[
\mathcal{L}_{\text{ans}} = -E_{q(F|E')} [\log p(A'|M, F)] + KL(q(F|E') \parallel p(F|E^*)) \\
= -E_{q(F|E')} [\log p(A'|M, F)] + \frac{1}{2} \log \left| \frac{\Sigma_{E'}}{\Sigma_{E^*}} \right| - d_f + tr\{\Sigma_{E'}^{-1}\Sigma_{E^*} - \Sigma_{E'}^{-1}\Delta\Sigma_{E'}\Delta\Sigma_{E'}^{-1}\}.
\]  

(3)

where we denote \( \Sigma_{E'} = \text{diag}(\sigma_{E'}^2), \Sigma_{E^*} = \text{diag}(\sigma_{E^*}^2), \Delta\mu = (\mu_{E'} - \mu_{E^*}) \) while \( d_f \) is the dimension of \( F \) and the detailed derivation of the KL divergence is included in Supplementary Material. Besides maximizing the marginal likelihood \( \log p(A'|M) \) of predicting true answer \( A' \), we also aim at maximizing the generative probability \( p(E'|M) \) of the ground truth explanation \( E' \) by the following loss:

\[
\mathcal{L}_{\text{exp}} = -\log p(E'|M).
\]  

(4)

By optimizing \( \mathcal{L}_{\text{ans}} \) and \( \mathcal{L}_{\text{exp}} \), we can train our SCM in Figure 3 (d) to predict accurate answers and generate rational explanations while modeling the causal correlation between explanation \( E \) and answer \( A \). We also prove the objectives proposed in this section and Section 4.1 are consistent in Supplementary Material. Next, we will introduce our specific reasoning model to implement the proposed SCM, of which the framework is shown in Figure 4.

### 4.3. Multimodal Content Encoder

To implement path \( Q \rightarrow M \leftarrow I \) in our SCM and compute \( M = M(Q, I) \), we adopt Vision-and-Language Pretrained Model (VLPM) \([25]\) (e.g., VisualBert \([20]\) and LXMERT \([42]\)) due to their promising ability of producing joint representations of vision and language. We follow REX \([11]\) to utilize pretrained Faster R-CNN \([40]\) to extract 36 visual objects \( \{f_i, p_i\}_{i=1}^{36} \) from image \( I \), where \( f_i \in \mathbb{R}^{2048} \) is the region-of-interest (ROI) feature and \( p_i \in \mathbb{R}^4 \) is the position vector of the \( i \)th object. For question words \( q_1, q_2, \ldots, q_m \), we add a \([CLS]\) token to the beginning and a \([EOS]\) token to the end (i.e., \( q_0 = [CLS], q_{m+1} = [EOS] \)). Then, we input all visual objects and question tokens into a VLPM to obtain the fused multimodal features:

\[
V, T = \text{VLPM}\{(f_i, p_i)\}_{i=1}^{36}, q_0 q_1 \cdots q_{m+1}),
\]  

(5)

where we adopt LXMERT as VLPM in experiments, \( V \in \mathbb{R}^{36 \times 768} \) are visual features of 36 image regions, \( T \in \mathbb{R}^{(m+2) \times 768} \) are question features of \( (m+2) \) question tokens, and we can obtain \( M = \{V, T\} \).
4.4. Multimodal Explanation Gating Transformer

To implement path $M \rightarrow E$ in our SCM and compute $E = E(M)$, different from the existing methods [11] that utilize LSTM-based decoders to generate the explanation, we propose Multimodal Explanation Gating Transformer (MEGT) to capture the complex cross-modal relationship. Our MEGT is based on Transformer [43]. However, different from traditional generative Transformers that generate tokens of a single modality, our model aims at generating both visual and textual tokens. Specially, we use token $\#j$ to represent the $j$th visual object. At the $t$th step of the explanation generation, we first obtain token embeddings of previous $(t-1)$ output tokens $\{e_i\}_{i=1}^{t}$ as follows:

$$emb(e_1) = \begin{cases} \text{wordemb}(e_i), & \text{if } e_i \text{ is a word token} \\ V_j, & \text{if } e_i = \#j \text{ is a visual token}, \end{cases} \quad (6)$$

where $\text{wordemb}(-)$ is a word embedding function. Denote $B = [emb(e_1), ..., emb(e_{t-1})] \in \mathbb{R}^{(t-1) \times 768}$, we model relationships among question words, visual regions, and generated explanation tokens to fuse comprehensive multimodal information by a $L$-layer Transformer as follows:

$$B^0 = B,$$
$$B^l = \text{TransLayer}^l(B^{l-1}), \quad (7)$$
$$B^l = \text{MultiHead}^l(B^l, [V, T], [V, T]),$$

where $\text{TransLayer}^l(\cdot)$ is the $l$th self-attention Transformer layer and $\text{MultiHead}^l(\cdot, \cdot, \cdot)$ is the $l$th multi-head attention layer proposed in [43].

Multimodal Gating Network

Inspired by [45], we utilize a gating function to determine whether generating a word token or a visual token at the $t$th step as follows:

$$\omega_t = \text{sigmoid}(LN(GELU(B^T_{l-1}, W^1_{g}))W^2_{g}) \in [0, 1], \quad (8)$$

where $W^1_g \in \mathbb{R}^{768 \times d_w}$ and $W^2_g \in \mathbb{R}^{d_w \times 1}$ are learnable matrices, $GELU(\cdot)$ is GELU function [14], and $LN(\cdot)$ is layer normalization function [6]. Then we adopt an MLP to predict the words in the vocabulary and utilize visual object features to predict the visual region numbers as follows:

$$y^w_t = \text{softmax}(LN(GELU(B^T_{l-1}, W^1_w))W^2_w) \in \mathbb{R}^U, \quad (9)$$

where $W^1_w \in \mathbb{R}^{T \times 768 \times d_w}$, $W^2_w \in \mathbb{R}^{d_w \times U}$ are learnable parameters of two linear layers, $B^T_{l-1}$ is the output feature of the $(t-1)$th token of the $l$th Transformer layer, and $U$ is the size of the vocabulary. By combining two probability vectors with the gating function, we obtain the final token probability as follows:

$$y^t_i = [\omega_i y^w_i (1 - \omega_i) y^v_i] \in \mathbb{R}^{U + 36}, \quad (10)$$

where $[\cdot | \cdot]$ is vector concatenation. During inference, the $t$th token is generated by taking the token with the highest prediction probability, i.e., $e^*_t = \text{argmax}_i y^t_i$.

4.5. Multimodal Explanation Encoder

To implement path $E \rightarrow F$ in our SCM and encode a robust explanation feature $F = F(E)$, we first insert a [CLS] token in the beginning of the explanation and utilize a Transformer to obtain contextual features as follows:

$$C = [emb([CLS]), emb(e_1), ..., emb(e_T)] \in \mathbb{R}^{(T+1) \times 768},$$

$$\bar{F} = \text{Transformer}(C) \in \mathbb{R}^{(T+1) \times 768}, \quad (11)$$
where $emb(\cdot)$ is the token embedding proposed in Equation 6, $T$ is the length of the explanation, and $\text{Transformer}(\cdot)$ is a 2-layer Transformer [43]. To obtain robust explanation feature, we assume $F \sim \mathcal{N}(\mu_E, \text{diag}(\sigma_E^2))$ and compute the parameters of the distribution as follows:

$$
\begin{bmatrix}
\mu_E \\
\log \sigma_E^2
\end{bmatrix} = \text{MLP}(\tilde{F}_0),
$$

(12)

where $\text{MLP}(\cdot)$ is a 2-layer MLP [33] and $\tilde{F}_0$ is the contextual feature of $[CLS]$. 

4.6. Answer Classifier

To implement path $F \rightarrow A \leftarrow M$ in our SCM and predict the answer $A = (A(M, F))$, we utilize explanation feature $F$ and multimodal context feature $T_0$ of $[CLS]$ token in Equation 5 and compute $p(A|M, F)$ as follows:

$$
p(A|M, F) = \text{softmax}(\text{LN}([F|T_0]|_a)) \in \mathbb{R}^K,
$$

(13)

where $[\cdot|\cdot]$ is concatenation function, $W_a \in \mathbb{R}^{1536 \times K}$ is a learnable matrix, and $K$ is the number of all possible answers. In the test, to avoid sampling bias and uncertain results, we compute the expectation of $p(A|M, F)$ to predict the answer by applying Normalized Weighted Geometric Mean (NWGM) approximation [7] as follows:

$$
E_F\{p(A|M, F)\} = E_F\{\text{softmax}(\text{LN}([F|T_0]|_a))\}
$$

$$
= \text{softmax}(\text{LN}([E|F|T_0]|_a)|_a) = \text{softmax}(\text{LN}([\mu_E|T_0]|_a|_a),
$$

(14)

where $E^*$ is the predicted explanation.

4.7. Optimization

We train our model by optimizing losses proposed in Equation 3 and 4 as follows:

$$
\mathcal{L}_{ans} = -\frac{1}{H} \sum_{i=1}^{H} A' \log p(A|M, F_i)
$$

$$
+ \frac{1}{2} \left[ \log \frac{\Sigma_{E'}}{\Sigma_{E'}} - d_f + \text{tr}\{\Sigma_{E'}^{-1}\Sigma_{E'}\} + \Delta \mu^T \Sigma_{E'}^{-1} \Delta \mu \right],
$$

$$
\mathcal{L}_{exp} = \sum_{t=1}^{T} e_t' \log y_t' + \sum_{i=1}^{T} [\omega_t' \log \omega_t + (1 - \omega_t') \log (1 - \omega_t)],
$$

$$
\mathcal{L} = \mathcal{L}_{ans} + \mathcal{L}_{exp},
$$

(15)

where $A'$ is the ground truth answer, $\omega_t$ is the ground truth gating value of the $t$th step, $e_t'$ is the $t$th ground truth explanation token, and we utilize Monte Carlo (MC) estimation to approximate the expectation as follows:

$$
-E_{q(F'|E')}\log p(A'|M, F) \approx -\frac{1}{H} \sum_{i=1}^{H} A' \log p(A|M, F_i),
$$

(16)

where $\{F_i \sim \mathcal{N}(\mu_{E'}, \text{diag}(\sigma_{E'}^2))\}_{i=1}^{H}$ are $H$ i.i.d. samples. Specially, $\mathcal{L}_{exp}$ aims at maximizing the prediction probabilities of both explanation tokens and gating values.

5. Experiments

In this paper, we focus on Explanatory Visual Question Answering (EVQA) task and conduct extensive experiments to verify the superiority of our VCIN. More details and results are included in Supplementary Material.

5.1. Datasets

We adopt the newly introduced GQA-REX [11] dataset, which expands upon the widely-used GQA [15] dataset by annotating multimodal explanations for visual reasoning processes. Specifically, GQA-REX is based on the balanced training set, balanced validation set, and standard test set of GQA. Moreover, we conduct experiments on GQA-OOD dataset [18], which has been recently introduced and contains out-of-distribution data.

5.2. Baseline Methods and Evaluation Metrics

To evaluate the effectiveness of the proposed VCIN method for EVQA, we compare it with three baseline approaches. VQAE [21] employs an LSTM-based language model for generating explanations and learns question answering jointly. EXP [45] utilizes an attention mechanism to integrate image features into an LSTM-based explanation generator. REX [11] is the state-of-the-art method, which employs a gating LSTM to generate explanations based on a fused input feature. The original REX (denoted as REX-VisualBert) employs VisualBert [20] as its backbone. To conduct a fair comparison with our VCIN that utilizes LXMERT, we also adopt a variant of REX (denoted as REX-LXMERT) that uses LXMERT as its backbone.

Following Chen and Zhao [11], we evaluate the model performance of visual question answering and multimodal explanation generation. To evaluate the visual question answering performance, we compute the answering Accuracy on validation and test sets. To evaluate the quality of the generated multimodal explanations, we employ five language metrics, namely BLEU-4 [30], METEOR [1], ROUGE-L [22], CIDEr [44], and SPICE [3]. Grounding metric [11] is utilized to evaluate the ability of correctly grounding visual regions in the generated explanations. Moreover, to evaluate the consistency between predicted answers and explanations, we propose a new automatic metric named Consistency (Con.) to compute the rate of explanations that contain the corresponding answers. More details about Con. are included in Supplementary Material. To align with human judgment, we also conduct a human evaluation. We design two criteria named Visual Consistency (Vis.) and Textual Consistency (Tex.) to evaluate
whether the predicted visual and textual tokens in explanations are consistent with the predicted answers, respectively. A 5-grade marking system is applied, with 5 as the maximum grade and 1 as the worst. We randomly select 500 validation samples and employ three professional annotators to conduct a blind evaluation. As annotated explanations are unavailable in the test set, we evaluate the generated explanations only on the validation set of GQA-REX.

### Table 2. Results of consistency between predicted answers and explanations on GQA-REX. The best results are highlighted in bold.

<table>
<thead>
<tr>
<th>Model</th>
<th>Con.</th>
<th>Vis.</th>
<th>Tex.</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>REX-VisualBert</td>
<td>74.69</td>
<td>2.82</td>
<td>3.77</td>
<td>3.30</td>
</tr>
<tr>
<td>REX-LXMERT</td>
<td>84.90</td>
<td>3.12</td>
<td>4.14</td>
<td>3.63</td>
</tr>
<tr>
<td>VCIN</td>
<td><strong>93.44</strong></td>
<td><strong>3.55</strong></td>
<td><strong>4.51</strong></td>
<td><strong>4.03</strong></td>
</tr>
</tbody>
</table>

### Table 3. Performance comparisons among variants of VCIN on GQA-REX. The best results are highlighted in bold.

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU-4</th>
<th>METEOR</th>
<th>ROUGE-L</th>
<th>CIDEr</th>
<th>SPICE</th>
<th>Grounding</th>
<th>GQA-val</th>
<th>GQA-test</th>
<th>OOD-val</th>
<th>OOD-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>VCIN-ANS</td>
<td>58.26</td>
<td>41.12</td>
<td>504.53</td>
<td>72.32</td>
<td>0.02</td>
<td>0.10</td>
<td>37.19</td>
<td>65.19</td>
<td>57.24</td>
<td>49.20</td>
</tr>
<tr>
<td>VCIN-EXP</td>
<td>57.19</td>
<td>40.89</td>
<td>513.82</td>
<td>74.25</td>
<td>0.02</td>
<td>0.00</td>
<td>36.92</td>
<td>64.19</td>
<td>57.16</td>
<td>49.50</td>
</tr>
<tr>
<td>VCIN-E2A</td>
<td>57.81</td>
<td>40.62</td>
<td>514.56</td>
<td>74.26</td>
<td>0.02</td>
<td>0.00</td>
<td>36.92</td>
<td>64.19</td>
<td>57.16</td>
<td>49.50</td>
</tr>
<tr>
<td>VCIN-RBF</td>
<td><strong>58.65</strong></td>
<td><strong>41.57</strong></td>
<td><strong>519.23</strong></td>
<td><strong>77.33</strong></td>
<td><strong>0.02</strong></td>
<td><strong>0.00</strong></td>
<td><strong>30.14</strong></td>
<td><strong>60.10</strong></td>
<td><strong>58.15</strong></td>
<td><strong>49.20</strong></td>
</tr>
<tr>
<td>VCIN</td>
<td><strong>94.44</strong></td>
<td><strong>43.55</strong></td>
<td><strong>54.15</strong></td>
<td><strong>87.33</strong></td>
<td><strong>0.02</strong></td>
<td><strong>0.00</strong></td>
<td><strong>37.19</strong></td>
<td><strong>70.19</strong></td>
<td><strong>78.15</strong></td>
<td><strong>58.15</strong></td>
</tr>
</tbody>
</table>

### 5.4. Ablation Study

To investigate the effectiveness of the proposed components, several variants are designed as follows: LININ-ANS abandons variational causal inference loss $L_{ans}$. LININ-EXP abandons explanation generation loss $L_{exp}$. LININ-E2A abandons the causal correlation from explanation to answer and predicts answers by $P(A|X)$. LININ-RBF is a causal variant that abandons the robust explanation feature $F$ and implement the joint model in Figure 3 (c).

We conduct experiments with the above variants on GQA-REX. The optimization procedure of all variants follows the proposed VCIN. In Table 3, the experimental results are listed, from which we have the following observations:

1. **VCIN significantly improves the quality of generated multimodal explanations.** Compared with REX-LXMERT, VCIN achieves relative improvements of 7.0%, 5.2%, 2.6%, 11.4%, 9.3%, and 9.2% for BLEU-4, METEOR, ROUGE-L, CIDEr, SPICE, and Grounding. This indicates that our multimodal explanation gating transformer can capture relations among visual regions, question words, and explanation tokens, leading to more coherent and rational explanations.

2. **VCIN significantly improves the accuracy of visual question answering.** While using the same backbone as REX-LXMERT, VCIN achieves answering accuracy improvements of 3.61%, 2.46%, 3.56%, and 2.14% on GQA-val, GQA-test, OOD-val, and OOD-test. This indicates that our proposed variational causal inference can effectively capture semantics in explanations and construct dependency between explanations and answers, resulting in more accurate answers.

(3) As shown in Table 2, both automatic metric and human evaluation show a significant improvement in the answer-explanation consistency of our proposed VCIN. Using the same backbone as REX-LXMERT, VCIN improves Cons. by 8.54% and relatively improves Vis. and Tex. by 13.8% and 8.9% respectively. These results verify that our proposed variational causal inference can effectively establish the consistency relation between the predicted answers and explanations to enhance the credibility of results.
5.5. Analysis of Causal Effects

To further investigate whether the proposed VCIN can learn the causal correlation between explanation and answer, we analyze the causal effects [31] of explanation $E$ on answer $A$ by manually intervening $E$ and observing the outcomes of $A$. In Figure 5, we demonstrate four examples in GQA-REX, from which we can find our VCIN changes the predicted answers to be consistent with the intervened explanations. For instance in Figure 5 (c), our VCIN predicts an explanation that the doll is wearing glasses and predicts the answer "glasses". After we manually replaced the word and visual object of "glasses" in the explanation with those of "shirt", VCIN changes its answer to "shirt" as well, which shows the causal dependency of $A$ on $E$. However, existing EVQA methods predict explanations and answers separately and intervening explanations cannot change their predicted answers. Those results indicate that our VCIN can learn the causal correlation between explanation and answer, based on which more consistent results can be inferred.

5.6. Qualitative Study

To further investigate the inferred results, we conduct a qualitative study on predicted answers and explanations. We show four examples in Figure 6 where REX uses LXMERT as the backbone for a fair comparison. Compared with the state-of-the-art REX, our proposed VCIN can perform better in terms of question answering, explanation generation, and answer-explanation consistency: (1) In (a) and (b), REX cannot ground the true objects in the images for explanation generation. However, our proposed multi-modal explanation gating transformer can capture the complex relationships among visual regions, question words, and explanation tokens. (2) (c) and (d) show the ability of our VCIN to accurately capture relations among various visual objects to infer explanations and answers, though the explanation in (c) is different from the ground truth. (3) In (b) and (c), the explanations and the answers inferred by REX are contradictory, while our VCIN can generate consistent results for all demonstrated examples. This improved answer-explanation consistency is important for a credible reasoning system.

5.7. Key Attributes in Explanation Generation

Following Chen and Zhao [11], we evaluate the ability of recognizing eight attributes (i.e., color, material, sport, shape, pose, size, activity, and relation) in explanations by calculating their recall rates on GQA-REX. To avoid trivial solutions, only questions where the attributes do not appear are considered. As shown in Table 4, VCIN significantly improves the recall rates of 8 key attributes related to different visual skills in the generated explanations. Compared with the state-of-the-art REX-LXMERT model, VCIN relatively improves Color, Material, Sport, Shape, Pose, Size, Activity, and Relation by 5.81%, 8.09%, 17.16%, 8.37%, 15.09%, 10.68%, 32.94%, and 8.48%, respectively. These results further demonstrate that our proposed VCIN can better capture diverse visual attributes to generate more coherent and rational explanations.

6. Conclusion

In this paper, we propose a novel Variational Causal Inference Network (VCIN) for explanatory visual question answering. To improve the consistency between the predicted answers and explanations, we propose a varia-
Table 4. Recall rates of key attributes related to different visual skills for explanation generation. The best results are highlighted in bold.

<table>
<thead>
<tr>
<th>Model</th>
<th>Color</th>
<th>Material</th>
<th>Sport</th>
<th>Shape</th>
<th>Pose</th>
<th>Size</th>
<th>Activity</th>
<th>Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>REX-VisualBert</td>
<td>56.01</td>
<td>49.27</td>
<td>72.77</td>
<td>40.64</td>
<td>74.80</td>
<td>65.31</td>
<td>46.58</td>
<td>29.00</td>
</tr>
<tr>
<td>REX-LXMERT</td>
<td>65.38</td>
<td>60.22</td>
<td>70.16</td>
<td>51.95</td>
<td>74.41</td>
<td>69.83</td>
<td>45.75</td>
<td>29.83</td>
</tr>
<tr>
<td>VCIN (Ours)</td>
<td><strong>69.18</strong></td>
<td><strong>65.09</strong></td>
<td><strong>82.20</strong></td>
<td><strong>56.30</strong></td>
<td><strong>85.64</strong></td>
<td><strong>77.29</strong></td>
<td><strong>60.82</strong></td>
<td><strong>32.36</strong></td>
</tr>
</tbody>
</table>

7. Acknowledgement

This work is supported by the National Key Research and Development Program of China (No.2021ZD0112202), the National Natural Science Foundation of China (No.62036012, 61936005, 61832002, 62276257, 61721004, 62072456, 62106262), and the Beijing Natural Science Foundation (L221004).

References


