Variational Causal Inference Network for Explanatory Visual Question Answering

Dizhan Xue[1,2] Shengsheng Qian[1,2] Changsheng Xu[1,2,3]
[1]State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[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 goes further by demanding an explanation of the reasoning process. This extension creates the possibility for improved explainability and credibility of reasoning models.

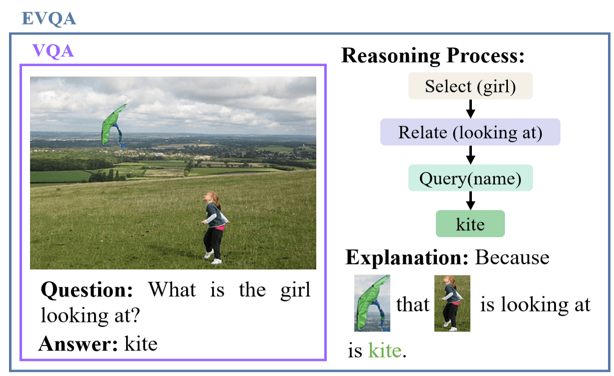
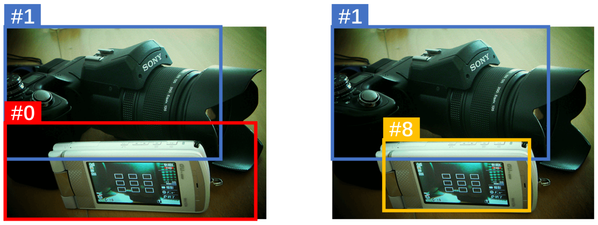


Figure 1. An example of Visual Question Answering (VQA) and Explanatory Visual Question Answering (EVQA): VQA requires answering the question with a related image, while EVQA additionally requires explaining the reasoning process.

Due to the definition of EVQA that the generated explanation should interpret the reasoning process of the inference model, it is crucial to maintain consistency between the predicted answer and explanation, or the inferred results can be incredible for users. However, existing methods for EVQA [11] predict the answer and the explanation separately based on the input multimodal information, which ignore the consistency relation between two outputs and may infer inconsistent results. As shown in Figure 2, the state-



Question: What is in front of the camera?
Answer: Microwave
Explanation: Because #0 in front of #1 is phone.

(a) REX [11]

Question: What is in front of the camera?
Answer: Phone
Explanation: Because #8 in front of #1 is phone.

(b) Ground Truth

Figure 2. An example of inconsistent answer and explanation inferred by REX on GQA-REX dataset.

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 can 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], ResNeXt [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 [17] are two typical language models. To fuse multimodal features, various vision-and-language pretrained models are proposed, such as VisualBERT [20], LXMERT [42], and Unicoder-VL [19]. Besides, some researchers design complex attention mechanisms to improve cross-modal interaction [53, 38, 28].

However, most VQA methods are based on DNNs and infer answers via black-box processes, which significantly reduces the explainability of results. Therefore, we focus on EVQA task, which aims at improving the explainability of VQA. REX [11] utilizes the fused input feature to generate the explanation via an LSTM-based generator. Differently, we propose a multimodal explanation gating transformer to capture complex relationships among question words, visual regions, and explanation tokens. Moreover, we propose a variational causal inference to improve the consistency between the predicted answers and explanations.

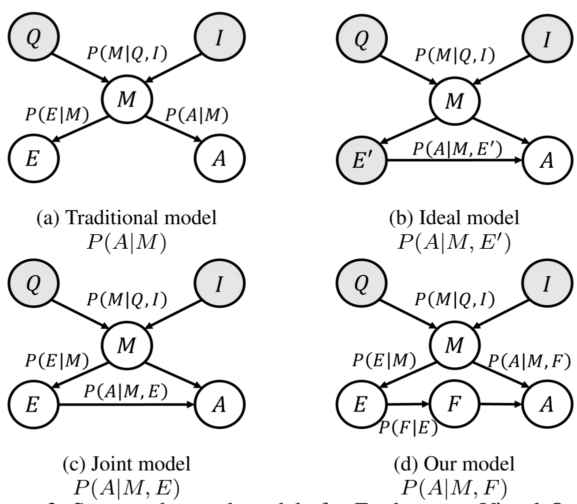


Figure 3. Structural causal models for Explanatory Visual Question Answering. Q denotes the textual question, I denotes the question-related image, M denotes multimodal content features, E and E' denote the explanation, F denotes robust explanation feature, and A denotes the answer. Gray nodes correspond 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). $Q = q_1 q_2 ... q_m$ is a textual question of total m words and q_i 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 $\{c_1, ..., c_K\}$ of total K answers. $E = e_1 e_2 ... e_n$ is the explanation of total n tokens, where e_i 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 $g : (Q, I) \mapsto (A, E)$, 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., $F \sim \mathcal{N}(\boldsymbol{\mu}_E, diag(\boldsymbol{\sigma}_E^2))$, where $\{\boldsymbol{\mu}_E, \boldsymbol{\sigma}_E^2\}$ 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 Evidence Lower Bound (ELBO) [49] of the marginal likelihood of predicting the true answer A' while modeling causal effects from explanation to answer as follows:

$$\begin{aligned}
&\log p(A'|M) \\
\geq &E_{q(F|M)}[\log p(A'|M, F) + \log p(F|M) - \log q(F|M)] \\
= &E_{q(F|M)}[\log p(A'|M, F)] - KL(q(F|M) \parallel p(F|M)) \\
= &E_{q(F|E')}[\log p(A'|M, F)] - KL(q(F|E') \parallel p(F|M)),
\end{aligned} \tag{1}$$

where we utilize the following lemma:

$$q(F|M) = \sum_E q(F|E)q(E|M) = q(F|E'), \tag{2}$$

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 $\{p(E|M)|\forall E\}$ 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:

$$\begin{aligned}
&\mathcal{L}_{ans} \\
= &- E_{q(F|E')}[\log p(A'|M, F)] + KL(q(F|E') \parallel p(F|E^*)) \\
= &- E_{q(F|E')}[\log p(A'|M, F)] \\
&+ \frac{1}{2}\Big[\log\frac{|\boldsymbol{\Sigma}_{E^*}|}{|\boldsymbol{\Sigma}_{E'}|} - d_f + tr\{\boldsymbol{\Sigma}_{E^*}^{-1}\boldsymbol{\Sigma}_{E'}\} + \Delta\boldsymbol{\mu}^T\boldsymbol{\Sigma}_{E^*}^{-1}\Delta\boldsymbol{\mu}\Big]
\end{aligned} \tag{3}$$

where we denote $\boldsymbol{\Sigma}_{E'} = diag(\boldsymbol{\sigma}_{E'}^2)$, $\boldsymbol{\Sigma}_{E^*} = diag(\boldsymbol{\sigma}_{E^*}^2)$, $\Delta\boldsymbol{\mu} = (\boldsymbol{\mu}_{E^*} - \boldsymbol{\mu}_{E'})$ 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:

$$\mathcal{L}_{exp} = -\log p(E'|M). \tag{4}$$

By optimizing $\mathcal{L}_{ans}$ and $\mathcal{L}_{exp}$, 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 $\{(\boldsymbol{f}_i, \boldsymbol{p}_i)\}_{i=1}^{36}$ from image I, where $\boldsymbol{f}_i \in \mathbb{R}^{2048}$ is the region-of-interest (ROI) feature and $\boldsymbol{p}_i \in \mathbb{R}^4$ is the position vector of the ith object. For question words $q_1q_2...q_m$, 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:

$$\boldsymbol{V}, \boldsymbol{T} = \mathrm{VLPM}(\{(\boldsymbol{f}_\mathrm{i}, \boldsymbol{p}_\mathrm{i})\}_{\mathrm{i}=1}^{36}, \mathrm{q_0q_1...q_{m+1}}), \tag{5}$$

where we adopt LXMERT as VLPM in experiments, $\boldsymbol{V} \in \mathbb{R}^{36\times768}$ are visual features of 36 image regions, $\boldsymbol{T} \in \mathbb{R}^{(m+2)\times768}$ are question features of $(m + 2)$ question tokens, and we can obtain $M = \{\boldsymbol{V}, \boldsymbol{T}\}$.

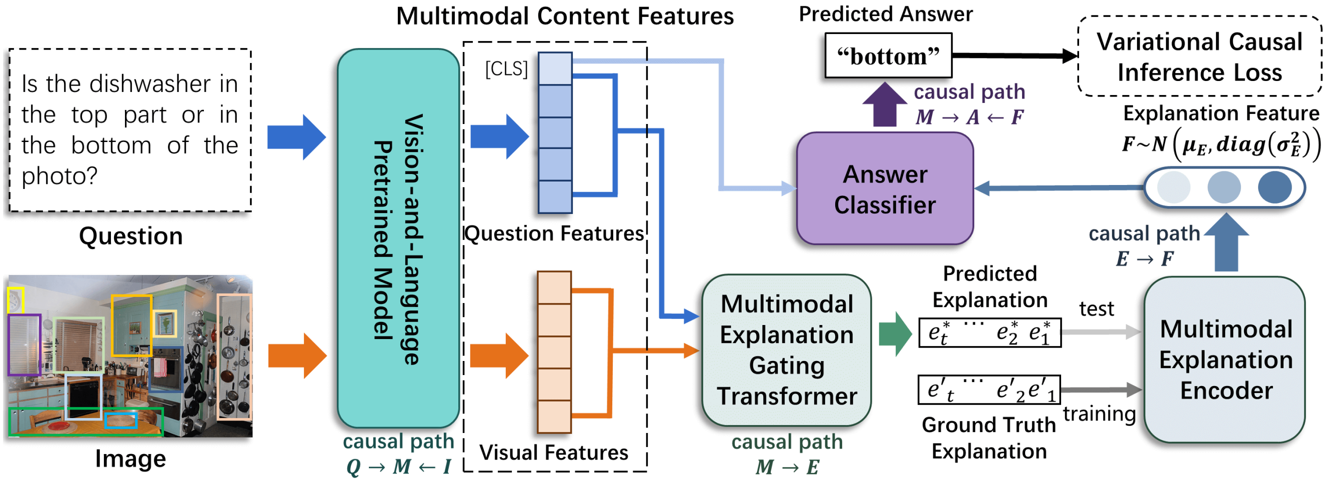


Figure 4. The framework of VCIN: (1) Causal path $Q \rightarrow M \leftarrow I$ is implemented by a Vision-and-Language Pretrained Model and multimodal content features are obtained; (2) Causal path $M \rightarrow E$ is implemented by the multimodal explanation gating transformer and multimodal explanation is predicted; (3) Causal path $E \rightarrow F$ is implemented by the multimodal explanation encoder and explanation feature is computed; (4) Causal path $M \rightarrow A \leftarrow F$ is implemented by the answer classifier and the answer is predicted.

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 jth visual object. At the tth step of the explanation generation, we first obtain token embeddings of previous $(t-1)$ output tokens $\{e_i\}_{i=1}^{t-1}$ as follows:

$$emb(e_i) = \begin{cases} wordemb(e_i), & \text{if } \mathrm{e_i} \text{ is a word token} \\ \boldsymbol{V}_j, & \text{if } \mathrm{e_i} = \#\mathrm{j} \text{ is a visual token,} \end{cases} \tag{6}$$

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

$$\begin{gathered} \boldsymbol{B}^0 = \boldsymbol{B}, \\ \bar{\boldsymbol{B}}^l = TransLayer^l(\boldsymbol{B}^{l-1}), \\ \boldsymbol{B}^l = MultiHead^l(\bar{\boldsymbol{B}}^l, [\boldsymbol{V}, \boldsymbol{T}], [\boldsymbol{V}, \boldsymbol{T}]), \\ l = 1, 2, ..., L, \end{gathered} \tag{7}$$

where $TransLayer^l(\cdot)$ is the lth self-attention Transformer layer and $MultiHead^l(\cdot,\cdot,\cdot)$ is the lth 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 tth step as follows:

$$\omega_t = sigmoid(LN(GELU(\boldsymbol{B}_{t-1}^L \boldsymbol{W}_g^1))\boldsymbol{W}_g^2) \in [0, 1], \tag{8}$$

where $\boldsymbol{W}_g^1 \in \mathbb{R}^{768\times d_g}$ and $\boldsymbol{W}_g^2 \in \mathbb{R}^{d_g \times 1}$ 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:

$$\begin{aligned} \boldsymbol{y}_t^w &= softmax(LN(GELU(\boldsymbol{B}_{t-1}^L \boldsymbol{W}_w^1))\mathbf{W}_w^2) \in \mathbb{R}^U, \\ \boldsymbol{y}_t^v &= softmax(\boldsymbol{B}_{t-1}^L \boldsymbol{V}^T) \in \mathbb{R}^{36}, \end{aligned} \tag{9}$$

where $\boldsymbol{W}_w^1 \in \mathbb{R}^{768\times d_o}$, $\boldsymbol{W}_w^2 \in \mathbb{R}^{d_o \times U}$ are learnable parameters of two linear layers, $\boldsymbol{B}_{t-1}^L$ is the output feature of the $(t-1)$th token of the Lth 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:

$$\boldsymbol{y}^t = [\omega_t \boldsymbol{y}_t^w | (1-\omega_t)\boldsymbol{y}_t^v] \in \mathbb{R}^{U+36}, \tag{10}$$

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

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:

$$\begin{gathered} \boldsymbol{C} = \left[emb([CLS]), emb(e_1), ..., emb(e_T)\right] \in \mathbb{R}^{(T+1)\times 768}, \\ \bar{\boldsymbol{F}} = Transformer(\boldsymbol{C}) \in \mathbb{R}^{(T+1)\times 768}, \end{gathered} \tag{11}$$

where $emb(\cdot)$ is the token embedding proposed in Equation 6, T is the length of the explanation, and $Transformer(\cdot)$ is a 2-layer Transformer [43]. To obtain robust explanation feature, we assume $F \sim \mathcal{N}(\boldsymbol{\mu}_E, diag(\boldsymbol{\sigma}_E^2))$ and compute the parameters of the distribution as follows:

$$\left[\begin{array}{c}\boldsymbol{\mu}_E \\ \log \boldsymbol{\sigma}_E^2\end{array}\right] = MLP(\bar{\boldsymbol{F}}_0), \tag{12}$$

where $MLP(\cdot)$ is a 2-layer MLP [33] and $\bar{\boldsymbol{F}}_0$ 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 $\boldsymbol{T}_0$ of $[CLS]$ token in Equation 5 and compute $p(A|M, F)$ as follows:

$$\begin{aligned} &p(A|M,F) \\ =&softmax(LN([F|\boldsymbol{T}_0])\boldsymbol{W}_a) \in \mathbb{R}^K, \end{aligned} \tag{13}$$

where $[\cdot|\cdot]$ is concatenation function, $\boldsymbol{W}_a \in \mathbb{R}^{1536 \times K}$ 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:

$$\begin{aligned} E_F\{p(A|M,F)\} =&E_F\{softmax(LN([F|\boldsymbol{T}_0])\boldsymbol{W}_a)\} \\ \approx&softmax(LN([E\{F\}|\boldsymbol{T}_0])\boldsymbol{W}_a) \\ =&softmax(LN([\boldsymbol{\mu}_{E^*}|\boldsymbol{T}_0])\boldsymbol{W}_a), \end{aligned} \tag{14}$$

where E^* is the predicted explanation.

4.7. Optimization

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

$$\begin{aligned} \mathcal{L}_{ans} =& -\frac{1}{H}\sum_{i=1}^{H} A' \log p(A|M, F_i) \\ &+\frac{1}{2}\Big[\log\frac{|\boldsymbol{\Sigma}_{E^*}|}{|\Sigma_{E'}|} - d_f + tr\{\boldsymbol{\Sigma}_{E^*}^{-1}\boldsymbol{\Sigma}_{E'}\} + \Delta\boldsymbol{\mu}^T\boldsymbol{\Sigma}_{E^*}^{-1}\Delta\boldsymbol{\mu}\Big], \\ \mathcal{L}_{exp} =& \sum_{t=1}^{T} e'_t \log \boldsymbol{y}^t + \sum_{t=1}^{T}[\omega'_t \log \omega_t + (1-\omega'_t)\log(1-\omega_t)], \\ \mathcal{L} =& \mathcal{L}_{ans} + \mathcal{L}_{exp}, \end{aligned} \tag{15}$$

where A' is the ground truth answer, ω'_t is the ground truth gating value of the tth step, e'_t is the tth ground truth explanation token, and we utilize Monte Carlo (MC) estimation to approximate the expectation as follows:

$$-E_{q(F|E')} \log p(A'|M, F) \approx -\frac{1}{H}\sum_{i=1}^{H} A' \log p(A|M, F_i), \tag{16}$$

where $\{F_i \sim \mathcal{N}(\boldsymbol{\mu}_{E'}, diag(\boldsymbol{\sigma}_{E'}^2))\}_{i=1}^H$ 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

Table 1. Results on explanation generation and question answering for Explanatory Visual Question Answering. GQA- and OOD- denote answering accuracy on GQA-REX and GQA-OOD datasets. The best results are highlighted in bold.

Model	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE	Grounding	GQA-val	GQA-test	OOD-val	OOD-test
VQAE	42.56	34.51	73.59	358.20	40.39	31.29	65.19	57.24	49.20	46.28
EXP	42.45	34.46	73.51	357.10	40.35	33.52	65.17	56.92	49.43	47.69
REX-VisualBert	54.59	39.22	78.56	464.20	46.80	67.95	66.16	57.77	50.26	48.26
REX-LXMERT	54.79	39.51	79.41	466.01	49.98	70.79	78.19	58.15	71.23	52.15
VCIN	**58.65**	**41.57**	**81.45**	**519.23**	**54.63**	**77.33**	**81.80**	**60.61**	**74.79**	**54.29**

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.

Model	Con.	Vis.	Tex.	Average
REX-VisualBert	74.69	2.82	3.77	3.30
REX-LXMERT	84.90	3.12	4.14	3.63
VCIN	**93.44**	**3.55**	**4.51**	**4.03**

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

Model	BLEU-4	METEOR	CIDEr	Grounding	GQA-val	Con.
VCIN-ANS	58.26	41.12	504.53	72.52	0.02	0.10
VCIN-EXP	0.03	6.79	0.00	22.52	80.58	0.00
VCIN-E2A	57.79	40.89	513.82	74.25	81.02	87.05
VCIN-RBF	57.81	40.62	514.56	74.26	78.73	90.56
VCIN	**58.65**	**41.57**	**519.23**	**77.33**	**81.80**	**93.44**

5.3. Results and Discussions

Table 1-2 shows the experimental results of all compared methods on GQA-REX and GQA-OOD. From the results, 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 Con. 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.

5.4. Ablation Study

To investigate the effectiveness of the proposed components, several variants are designed as follows: **LININ-ANS** abandons variational causal inference loss $\mathcal{L}_{ans}$. **LININ-EXP** abandons explanation generation loss $\mathcal{L}_{exp}$. **LININ-E2A** abandons the causal correlation from explanation to answer and predicts answers by $P(A|M)$. **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) The answering accuracy (GQA-val) of VCIN-ANS and the explanation quality (BlEU-4, METEOR, CIDEr, and Grounding) of VCIN-EXP sharply decrease. This shows the effectiveness of losses $\mathcal{L}_{ans}$ and $\mathcal{L}_{exp}$ for question answering and explanation generation. (2) All metric scores especially the answer-explanation consistency (Con.) of VCIN-E2A decrease. This is because VCIN-E2A abandons the causal correlation between explanations and answers. These results indicate that our variational causal inference can effectively model the causal correlation and improve the consistency. (3) All metric scores especially the answering accuracy of VCIN-RBF decline. This is due to the distribution shift between the ground truth explanations in training and the predicted explanations in test. These re-

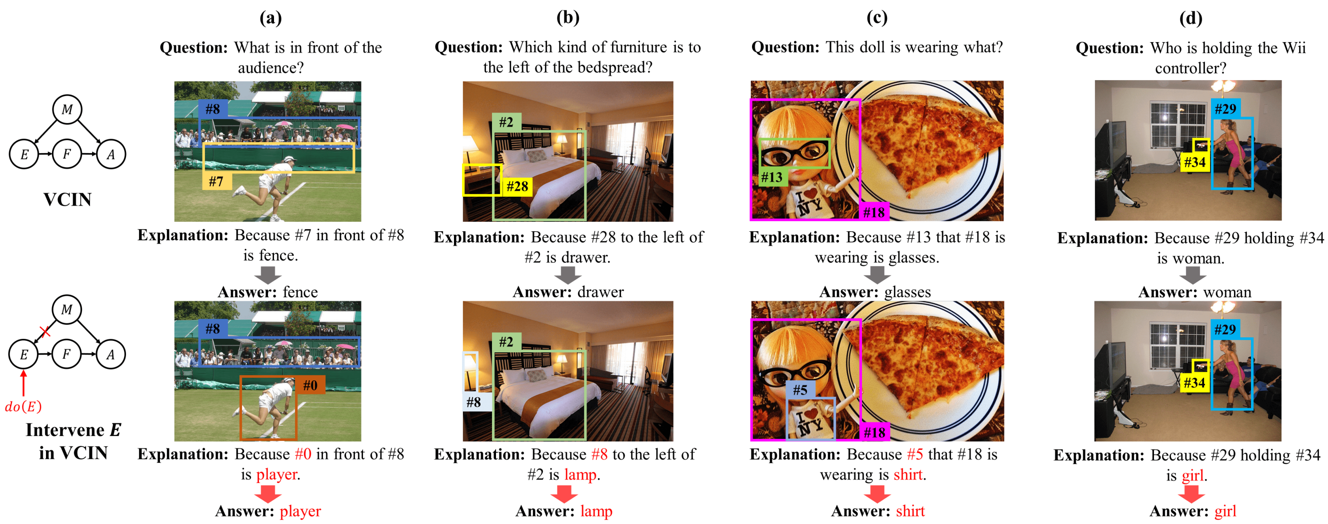


Figure 5. Causal results of manually intervening explanation E in our trained VCIN. Q and I are omitted in SCMs for brevity.

sults also verify that our proposed robust explanation feature can enhance the robustness of the causal correlation.

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-

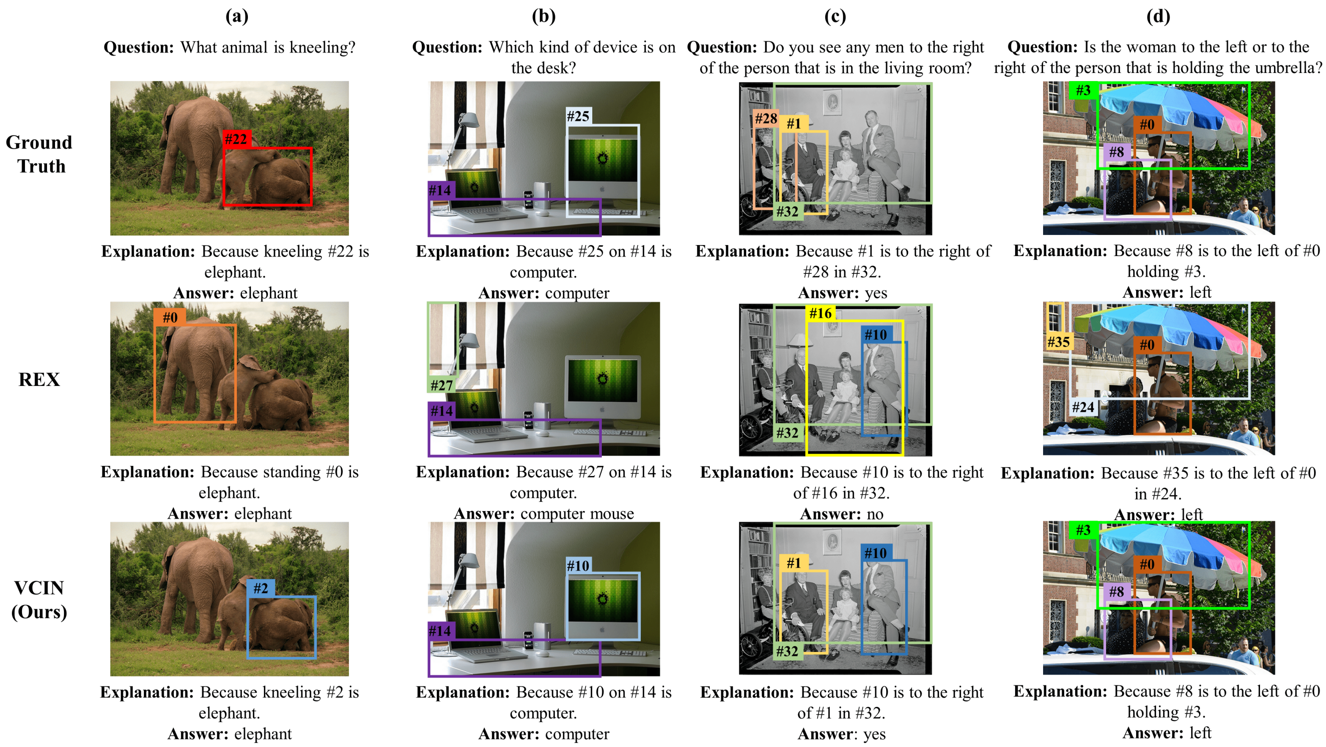


Figure 6. Qualitative results of different models for EVQA. Visual grounding is represented with the token #.

Table 4. Recall rates of key attributes related to different visual skills for explanation generation. The best results are highlighted in bold.

Model	Color	Material	Sport	Shape	Pose	Size	Activity	Relation
REX-VisualBert	56.01	49.27	72.77	40.64	74.80	65.31	46.58	29.00
REX-LXMERT	65.38	60.22	70.16	51.95	74.41	69.83	45.75	29.83
VCIN	**69.18**	**65.09**	**82.20**	**56.30**	**85.64**	**77.29**	**60.82**	**32.36**

tional causal inference to establish the causal correlation between the answer and explanation. To improve multimodal explanation generation, we design a multimodal explanation gating transformer to capture complex relationships among visual regions, question words, and explanation tokens. Extensive experiments indicate the superiority of VCIN in terms of answering accuracy, explanation quality, and answer-explanation consistency. In the future, we will attempt to apply the variational causal inference in more reasoning tasks to improve credibility and explainability.

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

[1] Abhaya Agarwal and Alon Lavie. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. *Proceedings of WMT-08*, 2007.

[2] Vedika Agarwal, Rakshith Shetty, and Mario Fritz. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9690–9698, 2020.

[3] 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.

[4] 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.

[5] 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.

[6] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[7] Pierre Baldi and Peter Sadowski. The dropout learning algorithm. *Artificial intelligence*, 210:78–122, 2014.

[8] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[9] Paola Cascante-Bonilla, Hui Wu, Letao Wang, Rogerio S Feris, and Vicente Ordonez. Simvqa: Exploring simulated environments for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5056–5066, 2022.

[10] Hongyu Chen, Ruifang Liu, and Bo Peng. Cross-modal relational reasoning network for visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3956–3965, 2021.

[11] Shi Chen and Qi Zhao. Rex: Reasoning-aware and grounded explanation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15586–15595, 2022.

[12] 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.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[15] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

[16] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. In defense of grid features for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10267–10276, 2020.

[17] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[18] Corentin Kervadec, Grigory Antipov, Moez Baccouche, and Christian Wolf. Roses are red, violets are blue... but should vqa expect them to? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2776–2785, 2021.

[19] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11336–11344, 2020.

[20] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[21] Qing Li, Qingyi Tao, Shafiq Joty, Jianfei Cai, and Jiebo Luo. Vqa-e: Explaining, elaborating, and enhancing your answers for visual questions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 552–567, 2018.

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

[23] Xiangru Lin, Yuyang Chen, Guanbin Li, and Yizhou Yu. A causal inference look at unsupervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1620–1629, 2022.

[24] Bing Liu, Dong Wang, Xu Yang, Yong Zhou, Rui Yao, Zhiwen Shao, and Jiaqi Zhao. Show, deconfound and tell: Image captioning with causal inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18041–18050, 2022.

[25] Siqu Long, Feiqi Cao, Soyeon Caren Han, and Haiqin Yang. Vision-and-language pretrained models: A survey. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5530–5537, 7 2022.

[26] David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Scholkopf, and Léon Bottou. Discovering causal signals in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6979–6987, 2017.

[27] Roxana Moreno and Richard Mayer. Interactive multimodal learning environments. *Educational psychology review*, 19(3):309–326, 2007.

[28] Binh X Nguyen, Tuong Do, Huy Tran, Erman Tjiputra, Quang D Tran, and Anh Nguyen. Coarse-to-fine reasoning for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4558–4566, 2022.

[29] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12700–12710, 2021.

[30] 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.

[31] Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96 – 146, 2009.

[32] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[33] Marius-Constantin Popescu, Valentina E Balas, Liliana Perescu-Popescu, and Nikos Mastorakis. Multilayer perceptron and neural networks. *WSEAS Transactions on Circuits and Systems*, 8(7):579–588, 2009.

[34] Shengsheng Qian, Hong Chen, Dizhan Xue, Quan Fang, and Changsheng Xu. Open-world social event classification. In *Proceedings of the ACM Web Conference 2023*, pages 1562–1571, 2023.

[35] Shengsheng Qian, Dizhan Xue, Quan Fang, and Changsheng Xu. Adaptive label-aware graph convolutional networks for cross-modal retrieval. *IEEE Transactions on Multimedia*, 24:3520–3532, 2021.

[36] Shengsheng Qian, Dizhan Xue, Quan Fang, and Changsheng Xu. Integrating multi-label contrastive learning with dual adversarial graph neural networks for cross-modal retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4794–4811, 2022.

[37] Shengsheng Qian, Dizhan Xue, Huaiwen Zhang, Quan Fang, and Changsheng Xu. Dual adversarial graph neural networks for multi-label cross-modal retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2440–2448, 2021.

[38] Tanzila Rahman, Shih-Han Chou, Leonid Sigal, and Giuseppe Carenini. An improved attention for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1653–1662, 2021.

[39] Dhanesh Ramachandram and Graham W Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine*, 34(6):96–108, 2017.

[40] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[41] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, 2019.

[42] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, 2019.

[43] 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.

[44] 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.

[45] Jialin Wu and Raymond Mooney. Faithful multimodal explanation for visual question answering. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 103–112, 2019.

[46] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[47] Dizhan Xue, Shengsheng Qian, Quan Fang, and Changsheng Xu. Mmt: Image-guided story ending generation with multimodal memory transformer. In *Proceedings of the 30th ACM International Conference on Multimedia*, page 750–758, 2022.

[48] Chenxiao Yang, Qitian Wu, Qingsong Wen, Zhiqiang Zhou, Liang Sun, and Junchi Yan. Towards out-of-distribution sequential event prediction: A causal treatment. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[49] Xitong Yang. Understanding the variational lower bound. *variational lower bound, ELBO, hard attention*, 22:1–4, 2017.

[50] Xiaofeng Yang, Guosheng Lin, Fengmao Lv, and Fayao Liu. Trrnet: Tiered relation reasoning for compositional visual question answering. In *European Conference on Computer Vision*, pages 414–430. Springer, 2020.

[51] Xu Yang, Hanwang Zhang, Guojun Qi, and Jianfei Cai. Causal attention for vision-language tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9847–9857, 2021.

[52] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xian-Sheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 33:655–666, 2020.

[53] Chen Zheng, Quan Guo, and Parisa Kordjamshidi. Cross-modality relevance for reasoning on language and vision. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7642–7651, 2020.