International Journal of Computer Vision (2024) 132:185–207
https://doi.org/10.1007/s11263-023-01858-y

Correlation Information Bottleneck: Towards Adapting Pretrained Multimodal Models for Robust Visual Question Answering

Jingjing Jiang[1] · **Ziyi Liu**[1] · **Nanning Zheng**[1]

Received: 3 January 2023 / Accepted: 18 July 2023 / Published online: 28 August 2023

Abstract

Benefiting from large-scale pretrained vision language models (VLMs), the performance of visual question answering (VQA) has approached human oracles. However, finetuning such models on limited data often suffers from overfitting and poor generalization issues, leading to a lack of model robustness. In this paper, we aim to improve input robustness from an information bottleneck perspective when adapting pretrained VLMs to the downstream VQA task. Input robustness refers to the ability of models to defend against visual and linguistic input variations, as well as shortcut learning involved in inputs. Generally, the representations obtained by pretrained VLMs inevitably contain irrelevant and redundant information for a specific downstream task, resulting in statistically spurious correlations and insensitivity to input variations. To encourage representations to converge to a minimal sufficient statistic in multimodal learning, we propose Correlation Information Bottleneck (CIB), which seeks a tradeoff between compression and redundancy in representations by minimizing the mutual information (MI) between inputs and representations while maximizing the MI between outputs and representations. Moreover, we derive a tight theoretical upper bound for the mutual information between multimodal inputs and representations, incorporating different internal correlations that guide models to learn more robust representations and facilitate modality alignment. Extensive experiments consistently demonstrate the effectiveness and superiority of the proposed CIB in terms of input robustness and accuracy.

Keywords Information bottleneck · Robustness · Visual question answering · Vision-language model

1 Introduction

Visual Question Answering (VQA) is a typical multimodal task that answers a given question based on image understanding (Antol et al., 2015). Recently, large-scale pretrained Vision-Language Models (VLMs) (Wang et al., 2023; Zeng et al., 2022; Wang et al., 2022; Yu et al., 2022; Wang et al., 2022; Li et al., 2022; Yuan et al., 2021; Wang et al., 2021) have advanced VQA performance to the level of human oracle. However, finetuning such pretrained VLMs on limited data for the downstream VQA task often leads to overfitting and poor generalization, limiting the improvement in robustness that pretrained VLMs can offer compared to the improvement in accuracy.

In this paper, we investigate how to effectively improve input robustness when adapting pretrained VLMs to a downstream VQA task. Input robustness in VQA refers to the ability of models to defend against visual variations (such as question-related object removal in images (Agarwal et al., 2020)), linguistic variations (such as word substitution and sentence rephrasing in questions (Shah et al., 2019)), and multimodal shortcut learning involved in input images and questions (Dancette et al., 2021). Practically, during the finetuning process, VQA is usually formulated as a multi-answer classification problem or a text generation problem, where pretrained multimodal transformers act as representation extractors with rich knowledge and are utilized to extract vision-language representations for answer prediction. As such, improving the input robustness of models essentially

Communicated by Jifeng Dai.

✉ Nanning Zheng
nnzheng@mail.xjtu.edu.cn

Jingjing Jiang
jingjingjiang2017@gmail.com

Ziyi Liu
liuziyi@stu.xjtu.edu.cn

[1] Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an 710049, Shaanxi, China

means obtaining more compact and task-related representations.

To this end, we propose to improve input robustness from an information-theoretical perspective. The representations yielded by pretrained VLMs inevitably contain irrelevant and redundant information for the specific downstream task, which is one possible reason for poor robustness. This is because irrelevant information encourages models to learn statistically spurious correlations between representations and labels, while task-agnostic redundant information reduces the sensitivity of models to input variations. Therefore, the two factors will compromise the input robustness of the model. To obtain more robust and compact representations, we thus anticipate that when adapting pretrained VLMs to VQA, these pretrained VLMs can discard irrelevant and redundant information in representations while preserving task-relevant information. The information bottleneck principle (Tishby et al., 2000) is adept at seeking a tradeoff between representation compression and redundancy. Motivated by this insight, we explore how to elegantly generalize the information bottleneck to find the minimal sufficient statistic for the learned representations, thereby improving the input robustness of VQA models.

We propose Correlation Information Bottleneck (CIB) to enhance input robustness when adapting pretrained VLMs to the downstream VQA task. Overall, by minimizing mutual information (MI) between representations and inputs while maximizing MI between representations and outputs, CIB seeks an optimal tradeoff between compression and redundancy in the representations learned by pretrained VLMs, enabling representations to converge to a minimal sufficient statistic. In detail, to accurately estimate the MI between multimodal inputs and representations, we derive a tight upper bound for the symmetrized joint MI, which measures different internal correlations rather than the overall dependency between different modalities. More specifically, the upper bound incorporates correlations between single-modal input and representation, as well as the correlation between visual and linguistic representations, guiding VQA models to learn more robust representations and better capture actual relationships. In particular, the multimodal representation correlation can facilitate modality alignment. Moreover, to ensure applicability to different transformer architectures, i.e., single-stream encoder, two-stream encoder, and encoder-decoder, we unify the internal representations of different pretrained VLMs using the representations after visual and linguistic embedding layers for CIB estimation.

To demonstrate the proposed CIB, we first provide rigorous theoretical proofs. Subsequently, using CIB as the training objective, we finetune pretrained VLMs including VisualBERT (Li et al., 2019), ViLBERT (Lu et al., 2019), VL-BERT$_{\mathrm{B}}$ (Su et al., 2020), VL-T5 (Cho et al., 2021), LXMERT (Tan & Bansal, 2019), UNITER$_{\mathrm{B}}$ (Chen et al., 2020), ALBEF (Li et al., 2021), mPLUG$_{\mathrm{B}}$ (Li et al., 2022), and BEiT-3$_{\mathrm{B}}$ (Wang et al., 2023) under a standard and clean data setting, and evaluate input robustness on five robustness benchmark datasets: VQA-Rephrasings (Shah et al., 2019), VQA P2 (Whitehead et al., 2020), IV-VQA (Agarwal et al., 2020), CV-VQA (Agarwal et al., 2020), and VQA-CE (Dancette et al., 2021). Extensive experiments and analyses consistently demonstrate that CIB significantly improves input robustness and exhibits advantages over existing methods when adapting pretrained VLMs to the downstream VQA task.

In summary, our main contributions are as follows: (*i*) We propose Correlation Information Bottleneck (CIB), a generic objective that can encourage representations to converge to a minimal sufficient statistic and enhance input robustness when adapting pretrained VLMs to VQA. (*ii*) We derive a tight upper bound for the MI between multimodal inputs and representations, incorporating different internal correlations that can guide models to learn more robust representations and facilitate modality alignment. (*iii*) Theoretical proofs and extensive experiments evaluate the robustness, superiority, and generalizability of our CIB.

The remainder of the paper is organized as follows: Sect. 2 introduces related literature on robustness in VQA, information bottleneck, and vision-language models. Section 3 elaborates on CIB, the application of CIB in adapting pretrained VLMs to VQA, and the theoretical analysis of input robustness for CIB. In Sect. 4, we conduct comprehensive experiments and discussions to demonstrate the effectiveness and superiority of CIB in terms of robustness and accuracy. In Appendix A, we provide a theoretical derivation for CIB and proofs for some proposed theorems.

2 Related Work

2.1 Robustness in VQA

Recently, in order to promote practical applications, numerous studies have been proposed to investigate various aspects of VQA robustness, such as input robustness (Shah et al., 2019; Whitehead et al., 2020; Agarwal et al., 2020; Kant et al., 2021), human-adversarial robustness (Li et al., 2021; Sheng et al., 2021), and robustness against answer distribution shift (Agrawal et al., 2022; Pan et al., 2022; Kervadec et al., 2021; Jiang et al., 2021; Teney et al., 2020; Clark et al., 2019; Goyal et al., 2017). In this paper, we explore input robustness, which refers to the capability of VQA models to defend against visual and linguistic variations, such as rephrasing questions (Shah et al., 2019; Whitehead et al., 2020), manipulating images (Agarwal et al., 2020), and shortcut learning involved in multimodal inputs (Dancette et al., 2021). The prevailing method to improve input robustness is data augmentation, i.e., generating additional data to train more robust VQA models. While data augmentation is

a feasible and effective solution, the quality of the generated data is uncontrollable (e.g., limited expressiveness and excessive verbosity), and the human-generated process is time-consuming. Moreover, cycle-consistency between the original question and its rephrasings (Shah et al., 2019), contrastive learning (Kant et al., 2021), and adversarial training (Li et al., 2020) have also been introduced to improve input robustness. These recent studies demonstrate that state-of-the-art VQA models remain vulnerable to input variation attacks. Therefore, in this paper, we focus on further improving the input robustness of existing VQA models.

2.2 Information Bottleneck

The Information Bottleneck (IB) principle was originally proposed by Tishby et al. (2000) for information compression, and was later applied to analyze deep learning model architectures (Tishby & Zaslavsky, 2015; Shwartz-Ziv & Tishby, 2017). Essentially, the IB objective is to seek a tradeoff between maximizing predictive accuracy and minimizing representation complexity. Some recent research targets exploiting the IB principle to improve model robustness and generalization, especially in domain generalization (Du et al., 2020; Li et al., 2022), out-of-distribution generalization (Ahuja et al., 2021), multiview representation learning (Federici et al., 2020; Bao, 2021), and finetuning of pretrained language models (Mahabadi et al., 2021; Wang et al., 2021; Dong et al., 2021). In addition, some works (Wang et al., 2022; Zhou et al., 2022; Pan et al., 2021; Jeon et al., 2021; Dubois et al., 2020) aim to learn disentangled optimal representations from an IB perspective. Since IB can facilitate compact and meaningful representation learning, we extend it to multimodal learning and apply IB to obtain robust VQA models.

2.3 Vision-Language Models

Vision-Language pretraining aims to learn task-agnostic visiolinguistic representations for improving the performance of downstream tasks in a finetuning fashion (Huang et al., 2020; Zhou et al., 2020; Shi et al., 2020; Li et al., 2021; Kim et al., 2021; Sun et al., 2021; Huang et al., 2021; Dou et al., 2022; Zhong et al., 2022; Alayrac et al., 2022; Xu et al., 2023). From the perspective of model architecture, prevailing pretrained vision-language models (VLMs) can be roughly grouped into three types: single-stream encoder (Su et al., 2020; Chen et al., 2020; Gan et al., 2020; Li et al., 2020; Zhang et al., 2021; Kim et al., 2021), two-stream encoder (Lu et al., 2019; Tan & Bansal, 2019; Lu et al., 2020; Yu et al., 2021; Li et al., 2021), and encoder-decoder (Cho et al., 2021; Li et al., 2021; Zeng et al., 2022; Li et al., 2022; Wang et al., 2022; Li et al., 2022). Specifically, single-stream models first align image regions and text tokens and then apply a uniform transformer (Vaswani et al., 2017) to learn the contextualized representations. Two-stream models first utilize two separate transformers to learn high-level representations for images and texts, and then integrate the two modalities with a cross-modal transformer. Encoder-decoder models respectively utilize encoders and decoders to learn multimodal representations and to generate related texts for specific downstream tasks. In this paper, we unify the three typical types of VLMs and propose CIB to improve input robustness when adapting these pretrained VLMs for the downstream VQA task.

3 Methodology

In this section, we first present the preliminaries of the problem setting and the general IB principle. Then, we elaborate on the proposed CIB in Sect. 3.2 and explain how to apply CIB to improve input robustness when adapting pretrained VLMs to VQA in Sect. 3.3.

3.1 Preliminary

Problem Setting. In the finetuning process, single-stream and two-stream VLMs usually formulate the VQA task as a multi-answer classification problem Chen et al. (2020), Tan and Bansal (2019), while encoder-decoder VLMs often regard VQA as text generation (Cho et al., 2021; Wang et al., 2022), i.e., generating free-form textual answers for a given question instead of selecting a specific one from the predefined set of answers. Given a VQA dataset $\mathcal{D} = \{(I, Q, y) \in \mathcal{I} \times \mathcal{Q} \times \mathcal{Y}\}$, where I is an image, Q is a question, and y is an answer, VLMs take image-question pairs as input, where the image is further represented as a set of image regions or patches $\{v_1, \ldots, v_K\}$ (K is the number of regions or patches in one image) and the question is tokenized as a token sequence $\{w_1, \ldots, w_L\}$ (L is the number of word tokens in a question). For single-stream and two-stream VLMs, they output the answer probability distribution Y using an additional VQA Head module, which is implemented by two fully-connected layers sandwiched with GeLU activation and Layer Normalization operation. Meanwhile, encoder-decoder VLMs directly generate textual answers without any additional module.

IB View of Representation Learning. From an information-theoretic perspective, seeking a robust representation T in representation learning is equivalent to preserving information about the output Y while removing irrelevant and redundant information from the input X. This is because for a given task, irrelevant and redundant information may encourage models to learn superfluous correlations between answer labels and inputs. Formally, the IB principle (Tishby et al., 2000; Tishby & Zaslavsky, 2015) formulates representation

learning as an information tradeoff and finds an optimal representation by maximizing the Lagrangian

$$\mathscr{L}_{\text{IB}} := I(Y;T) - \beta I(X;T), \tag{1}$$

where $\beta \geq 0$ controls the tradeoff between compression and prediction, and $I(\cdot;\cdot)$ denotes mutual information (MI).

3.2 Correlation Information Bottleneck

In vision-language representation learning, given two modality inputs X^v and X^l, VLMs learn the corresponding visual and linguistic representations T^v and T^l of some intermediate transformer layers while simultaneously maximizing the MI between the obtained representations and a given label Y to guarantee representations contain sufficient information for predicting Y. To extend the general IB principle to the multimodal setting, we first consider the inputs and internal representations as a whole, i.e., $X = [X^v, X^l]$ and $T = [T^v, T^l]$, respectively, and then derive a differentiable estimation for IB by expanding the MI terms in Eq. (1).

Specifically, we first focus on $I(Y;T)$, which can be rewritten using the conditional probability definition:

$$I(Y;T) = \int p(y,t) \log \frac{p(y|t)}{p(y)} \, dydt. \tag{2}$$

Since the conditional probability $p(y|t)$ is intractable, we instead estimate $I(Y;T)$ with the BA (Barber & Agakov, 2003) lower bound:

$$I(Y;T) \geq \int p(y,t) \log q(y|t) dydt - \int p(y) \log p(y) dy, \tag{3}$$

where $q(y|t)$ is an accessible auxiliary distribution for $p(y|t)$ and $-\int p(y) \log p(y) dy = H(Y)$ is the entropy of labels, which is independent of the optimization procedure in finetuning. Ignoring $H(Y)$, the remaining term of the lower bound in Eq. (3) is equal to $-H(Y|T)$, meaning that maximizing the lower bound of $I(Y;T)$ is equivalent to minimizing the cross-entropy loss of a specific task. In other words, when using IB as the training objective, maximizing $I(Y;T)$ can be equivalent to minimizing the VQA loss $\mathscr{L}_{\text{vqa}}$.

Next, we consider the mutual information between the input sources and their corresponding representations, that is, the term $I(X;T)$ in Eq. (1). To accurately estimate $I(X;T)$, instead of directly measuring the overall dependency between X and T (i.e., regarding X^v and X^l as a whole one X, and regarding T^v and T^l as a whole one T), we consider expanding $I(X;T)$ to $I(X^v, X^l; T^v, T^l)$, and attempt to derive a tight upper bound for it. Since $I(X^v, X^l; T^v, T^l)$ incorporates different internal correlations, such as the correlation between visual input X^v and representation T^v, the correlation between linguistic input X^l and representation T^l, and the correlation between visual and linguistic representations (T^v and T^l). These correlations may guide models to learn more compact visual and linguistic representations and facilitate modality alignment between visual and linguistic representations. Therefore, we propose to maximize the Correlation Information Bottleneck (CIB) formula:

$$\mathscr{L}_{\text{CIB}} := I(Y;T) - \beta I(X^v, X^l; T^v, T^l), \tag{4}$$

where $I(X^v, X^l; T^v, T^l)$ is a symmetrized variant of joint mutual information (Bennasar et al., 2015) that considers the internal correlations between $X = [X^v, X^l]$ and $T = [T^v, T^l]$. To efficiently estimate $I(X^v, X^l; T^v, T^l)$, we first further expand it conditioned on the properties of mutual information and the data processing inequality in representation learning (Federici et al., 2020). The derivation can be formally stated by Theorem 1 (cf. Sect. 1 for proof):

Theorem 1 *(Upper Bound of $I(X^v, X^l; T^v, T^l)$) Given two groups of random variables $X = [X^v, X^l]$ and $T = [T^v, T^l]$, the MI $I(X^v, X^l; T^v, T^l)$ can be upper-bounded with*

$$\begin{aligned} I(X;T) &= I(X^v, X^l; T^v, T^l), \\ &\leq I(X^v;T^v) + I(X^l;T^l) - I(T^v;T^l) + D_{skl}, \end{aligned} \tag{5}$$

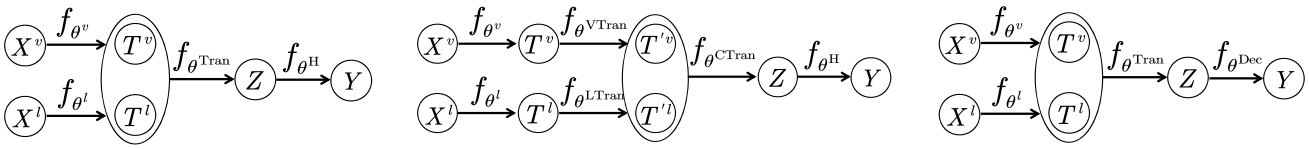
where D_{skl} denotes the symmetric Kullback–Leibler (KL) divergence that can be calculated by averaging the divergences $KL(p(t^v|x^v)||p(t^l|x^l))$ and $KL(p(t^l|x^l)||p(t^v|x^v))$.

After approximating the MI $I(X^v, X^l; T^v, T^l)$, the lower bound of $\mathscr{L}_{\text{CIB}}$ can be stated as the following Theorem 2.

Theorem 2 *(Lower Bound of CIB) Given random variable $X = [X^v, X^l]$, two deterministic functions f_{θ^v} and f_{θ^l} let $T^v = f_{\theta^v}(X^v)$ and $T^l = f_{\theta^l}(X^l)$. Correlation Information Bottleneck (CIB) can then be bounded as*

$$\begin{aligned} \mathscr{L}_{CIB} &= I(Y;T) - \beta I(X^v, X^l; T^v, T^l), \\ &\geq I(Y;T) - \beta \Big[I(X^v;T^v) + I(X^l;T^l) - I(T^v;T^l) + D_{skl} \Big]. \end{aligned} \tag{6}$$

In summary, Theorem 2 suggests that in vision-language representation learning, if $I(Y;T)$ is considered a task-related objective, $I(X^v, X^l; T^v, T^l)$ can be viewed as a regularizer used to constrain the compactness and redundancy of the learned representations. Overall, CIB encourages pretrained VLMs to learn more robust representations by seeking an optimal tradeoff between redundancy and compression in representations. Moreover, CIB facilitates modality alignment and correlation by maximizing the MI $I(T^v; T^l)$ between visual and linguistic representations.



(a) Single-Stream Encoder (b) Two-Stream Encoder (c) Encoder-Decoder

Fig. 1 The information flow of three typical transformer architectures of VLMs

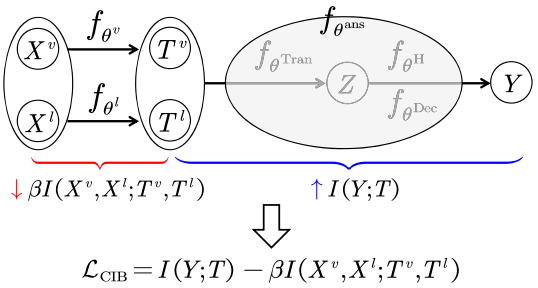


Fig. 2 Illustration of using CIB to adapt pretrained VLMs to downstream task. CIB seeks a minimal sufficient statistic by minimizing MI between input and internal representation (↓) while maximizing MI between output and representation (↑)

3.3 Adapting Pretrained VLMs to VQA with CIB

As illustrated in Fig. 1a, b, and c, there are three typical transformer architectures for VLMs: single-stream encoder (Li et al., 2019; Su et al., 2020; Chen et al., 2020), two-stream encoder (Tan & Bansal, 2019; Lu et al., 2019), and encoder-decoder (Cho et al., 2021; Li et al., 2021). When finetuning pretrained VLMs with CIB, to unify the three architectures into a single formulation, as shown in Fig. 2, we utilize the region-level or patch-level visual features after the visual Embedding layer (i.e., f_{θ^v} is the parametric embedding layer) as the internal visual representation T^v. Analogously, the token-level linguistic features after the linguistic embedding layer (f_{θ^l}) are considered as the internal linguistic representation T^l. All subsequent Transformer layers ($f_{\theta^{\text{Tran}}}$) and the VQA Head module ($f_{\theta^{\text{H}}}$) for the single-stream and two-stream VLMs as well as the Decoder ($f_{\theta^{\text{Dec}}}$) for the encoder-decoder VLMs serve as the parametric approximator ($f_{\theta^{\text{ans}}}$) to generate Y given $T = [T^v, T^l]$. As summarized in Algorithm 1, we first convert $I(Y; T)$ to the cross-entropy loss ($\mathscr{L}_{\text{vqa}}$) for answer prediction in VQA and estimate the remaining terms in Theorem 2. After obtaining $\mathscr{L}_{\text{CIB}}$, we update all parameters by minimizing $-\mathscr{L}_{\text{CIB}}$. Next, we elaborate on the estimation of CIB terms.

3.3.1 Estimating CIB Terms

As stated in Theorem 2, in addition to the task-related MI term $I(Y; T)$, $I(X^v, X^l; T^v, T^l)$ can be further decomposed into four computable MI terms. Firstly, we focus on the MI between inputs and representations within a single visual or linguistic modality. The inputs X^v and X^l are intrinsically two sets of random variables, i.e., $X^v = [X^v_1, ..., X^v_K]$ and $X^l = [X^l_1, ..., X^l_L]$. The functions f_{θ^v} and f_{θ^l} transform X^v and X^l into visual and linguistic representations, respectively, such that $T^v = [T^v_1, ..., T^v_K] = [f_{\theta^v}(X^v_1), ..., f_{\theta^v}(X^v_K)]$ and $T^l = [T^l_1, ..., T^l_L] = [f_{\theta^l}(X^l_1), ..., f_{\theta^l}(X^l_L)]$. While for sample pairs $\{(X^v_i, T^v_i)\}^K_{i=1}$ and $\{(X^l_i, T^l_i)\}^L_{i=1}$, the conditional probability distributions $p(t^v|x^v)$ and $p(t^l|x^l)$ are known during the finetuning process. Consequently, we adopt a sample-based differentiable MI estimator, CLUB (Cheng et al., 2020), to approximate the upper bound of the MI between visual or linguistic inputs and their corresponding representations, i.e.,

$$\hat{I}(X^v; T^v) = \frac{1}{K^2}\sum_{i=1}^{K}\sum_{j=1}^{K}\Big[\log p(t^v_i|x^v_i) - \log p(t^v_j|x^v_i)\Big], \tag{7}$$

$$\hat{I}(X^l; T^l) = \frac{1}{L^2}\sum_{i=1}^{L}\sum_{j=1}^{L}\Big[\log p(t^l_i|x^l_i) - \log p(t^l_j|x^l_i)\Big]. \tag{8}$$

Algorithm 1 Finetuning pretrained VLMs with CIB for VQA

Input: Visual (Image) sequence: X^v; Hyperparameter: β;
Linguistic (Question) sequence: X^l.
Output: Training loss: $-\mathcal{L}_{\text{CIB}}$.

Load pretrained weights for VLMs;
$T^v \leftarrow f_{\theta^v}(X^v)$, $T^l \leftarrow f_{\theta^l}(X^l)$;
$Y \leftarrow f_{\theta^{\text{ans}}}([T^v, T^l])$;
procedure ESTIMATE $\mathcal{L}_{\text{CIB}}$
  # estimate all terms in $\mathcal{L}_{CIB}$
  $\mathcal{L}_{\text{vqa}} \leftarrow$ convert $I(Y; T)$ to VQA loss;
  $\hat{I}(X^v; T^v) \leftarrow$ estimate MI of X^v and T^v with CLUB;
  $\hat{I}(X^l; T^l) \leftarrow$ estimate MI of X^l and T^l with CLUB;
  $\hat{I}(\bar{T}^v; \bar{T}^l) \leftarrow$ estimate MI of T^v and T^l using NWJ with $f_{\theta_{\text{FC}}}$;
  $\hat{D}_{\text{skl}} \leftarrow$ estimate symmetric KL between $p(t^l|x^l)$ and $p(t^v|x^v)$;
  # compute $\mathcal{L}_{CIB}$
  $\mathcal{L}_{\text{CIB}} = -\mathcal{L}_{\text{vqa}} - \beta\Big[\hat{I}(X^v; T^v) + \hat{I}(X^l; T^l) - \hat{I}(\bar{T}^v; \bar{T}^l) + \hat{D}_{\text{skl}}\Big]$;
Update f_{θ^v}, f_{θ^l}, $f_{\theta^{\text{ans}}}$, $f_{\theta_{\text{FC}}}$ by minimizing $-\mathcal{L}_{\text{CIB}}$.

For $I(T^v; T^l)$, it is challenging to estimate directly due to the different sequence lengths of $T^v \in \mathbb{R}^{K\times d}$ and $T^l \in \mathbb{R}^{L\times d}$. Therefore, we transform the two sequence representations into the global visual and linguistic representations

$\bar{T}^v \in \mathbb{R}^d$ and $\bar{T}^l \in \mathbb{R}^d$, using a one-layer fully-connected (FC) network. To guarantee that the inequality in Eq. (6) holds, we should approximate the lower bound of $I(T^v; T^l)$. Therefore, we estimate $I(T^v; T^l)$ with NWJ (Poole et al., 2019), i.e.,

$$\hat{I}(\bar{T}^v; \bar{T}^l) = \mathbb{E}_{p(\bar{t}^v, \bar{t}^l)}\left[\log f_{\theta_{\text{FC}}}(\bar{t}^v, \bar{t}^l)\right] - \frac{1}{e}\mathbb{E}_{p(\bar{t}^v)p(\bar{t}^l)}\left[f_{\theta_{\text{FC}}}(\bar{t}^v, \bar{t}^l)\right], \tag{9}$$

where $f_{\theta_{\text{FC}}}$ denotes the discriminant function implemented using a two-layer FC network.

Finally, for D_{skl} in $\mathscr{L}_{\text{CIB}}$, since $p(t^l|x^l)$ and $p(t^v|x^v)$ have a known probability density, we can directly compute the two KL divergences using internal visual and linguistic representations. That is, D_{skl} can be obtained by

$$\hat{D}_{\text{skl}} = \frac{1}{2}\left[\text{KL}\left(p(t^v|x^v)||p(t^l|x^l)\right) + \text{KL}\left(p(t^l|x^l)||p(t^v|x^v)\right)\right]. \tag{10}$$

3.3.2 Theoretical Justification for Input Robustness

In the following section, we conduct a theoretical analysis of input robustness for CIB. Formally, for a perturbation δ added to visual and linguistic inputs, let $X' = [X^{v\prime}, X^{l\prime}]$ represent the perturbed inputs of standard inputs $X = [X^v, X^l]$, i.e., $X' = X + \delta$. Functions f_{θ^v} and f_{θ^l} transform $X = [X^v, X^l]$ and $X' = [X^{v\prime}, X^{l\prime}]$ into $T = [T^v, T^l] = [f_{\theta^v}(X^v), f_{\theta^l}(X^l)]$ and $T' = [T^{v\prime}, T^{l\prime}] = [f_{\theta^v}(X^{v\prime}), f_{\theta^l}(X^{l\prime})]$, with $T \neq T'$. The distributions of X and X' are denoted by probabilities $p(x)$ and $q(x)$, where $q(x)$ approximates the distribution of $p(x)$. δ_m is the maximum perturbation bound that does not alter the output label, i.e., $Y = f_{\theta^{\text{ans}}}(T) = f_{\theta^{\text{ans}}}(T')$ when $||\delta|| \leq \delta_m$. According to the definition of CIB, the performance gap between standard inputs and perturbed inputs is $|I(T; Y) - I(T'; Y)| = |I(T^v, T^l; Y) - I(T^{v\prime}, T^{l\prime}; Y)|$. To provide theoretical justification for the performance gap, based on the work (Wang et al., 2021), we derive the upper bound

$$\begin{aligned}
&|I(T; Y) - I(T'; Y)| \\
&= |I(T^v, T^l; Y) - I(T^{v\prime}, T^{l\prime}; Y)|, \\
&\leq B_1^v\sqrt{\mathcal{T}^v}\left(I(X^v; T^v)\right)^{1/2} + B_2^v|\mathcal{T}^v|^{3/4}\left(I(X^v; T^v)\right)^{1/4} \\
&\quad + B_3^v\sqrt{|\mathcal{T}^v|}\left(I(X^{v\prime}; T^{v\prime})\right)^{1/2} + B_4^v|\mathcal{T}^v|^{3/4}\left(I(X^{v\prime}; T^{v\prime})\right)^{1/4} \\
&\quad + B_1^l\sqrt{\mathcal{T}^l}\left(I(X^l; T^l)\right)^{1/2} + B_2^l|\mathcal{T}^l|^{3/4}\left(I(X^l; T^l)\right)^{1/4} \\
&\quad + B_3^l\sqrt{|\mathcal{T}^l|}\left(I(X^{l\prime}; T^{l\prime})\right)^{1/2} + B_4^l|\mathcal{T}^l|^{3/4}\left(I(X^{l\prime}; T^{l\prime})\right)^{1/4} \\
&\quad + B_0^v + B_0^l,
\end{aligned} \tag{11}$$

where $\mathcal{T}^v$ is the finite support of T^v and $T^{v\prime}$, and B_0^v, B_1^v, B_2^v, B_3^v, and B_4^v are constants that depend on the sequence length K, δ, and $p(x^v)$. $\mathcal{T}^l$ is the finite support of T^l and $T^{l\prime}$, and B_0^l, B_1^l, B_2^l, B_3^l, and B_4^l are constants that depend on the sequence length L, δ, and $p(x^l)$ (cf. Sect. 1 for proof).

4 Experiment

In this section, we evaluate the input robustness of the proposed CIB and carry out detailed ablation studies to analyze the performance contribution of CIB components. Meanwhile, we explore the effectiveness of CIB in some other cases, such as standard VQA performance, adversarial attacks, and other multimodal tasks beyond VQA.

4.1 Experimental Settings

4.1.1 Evaluation Datasets

Unless otherwise specified, we finetune pretrained VLMs on the standard and clean VQA v2 training set (Goyal et al., 2017) and evaluate input robustness on five robustness benchmark datasets: VQA-Rephrasings (Shah et al., 2019), VQA P2 (Whitehead et al., 2020), IV-VQA (Agarwal et al., 2020), CV-VQA (Agarwal et al., 2020), and VQA-CE (Dancette et al., 2021). VQA-Rephrasings and VQA P2 evaluate robustness against linguistic variations, while IV-VQA and CV-VQA evaluate robustness against visual variations. VQA-CE, on the other hand, assesses robustness against shortcut learning involving inputs. As all these datasets are built on the VQA v2 (Goyal et al., 2017) validation split, *we consequently train our models only on the VQA v2 training set*.

Table 1 summarizes dataset details, including the type of perturbation, specific evaluation metrics for robustness, question type (QType), shared dataset for finetuning, and the test datasets statistics. These statistics encompass the total number of image-question pairs (#IQ), perturbation samples (#PER/CE), original samples (#ORI/Easy), and the average question length (len(Q)). Specifically, VQA-Rephrasings averagely collects 3 rephrasings for each of the 40,504 questions sampled from the VQA v2 validation set, resulting in approximately 162k image-question pairs. VQA P2 creates three types of linguistic perturbations, i.e., sentence rephrasing (Par), and word substitution with synonyms (Syn) or antonyms (Ant), for 25,814 sampled questions, ultimately obtaining roughly about 52k image-question pairs. IV-VQA employs a GAN-based resynthesis technique to remove objects irrelevant to the given question from the image, such that object removal does not affect the answer. Conversely, CV-VQA focuses on counting questions (Num) and removes one relevant object, causing the predicted answer

Table 1 Details on input robustness datasets

Datasets	Perturbation	Metric	QType	Finetuning Dataset	Evaluation			
					len(Q)	#IQ	#PER/CE	#ORI/Easy
VQA-Rephrasings (Shah et al., 2019)	Rephrasing	CS(m)	All	VQA v2 train	7.15	162k	121,516	40,504
VQA P2 (Whitehead et al., 2020)	Par&Syn&Ant	CS(m)	All	VQA v2 train	6.32	52k	26,512	25,814
IV-VQA (Agarwal et al., 2020)	Invariant object	#flips	All	VQA v2 train	5.85	120k	83,700	36,181
CV-VQA (Agarwal et al., 2020)	Covariant object	#flips	Num	VQA v2 train	5.83	4k	4,141	2,641
VQA-CE (Dancette et al., 2021)	Counterexample	–	All	VQA v2 train	6.19	214k	63,298	147,681

on the number of objects to be reduced by one. In total, IV-VQA and CV-VQA contain approximately 120k and 4k image-question pairs, respectively. VQA-CE is an evaluation benchmark for multimodal shortcuts involved in images and questions. It utilizes the detected shortcuts from the training set to obtain 63,298 counterexamples, where all shortcuts lead to incorrect answers, from the VQA v2 validation set. Additionally, VQA-CE constructs 147,681 easy examples in which at least one shortcut provides the correct answer.

Moreover, to analyze the effectiveness of CIB on standard VQA performance, we conduct experiments on VQA v2 (Goyal et al., 2017). Specifically, we first utilize CIB as the training objective to finetune pretrained VLMs on the VQA v2 training and validation sets, and subsequently test standard VQA performance on VQA v2 test-dev. To evaluate the generalizability of CIB to other multimodal tasks, we perform experiments on RefCOCO+ (Yu et al., 2016) in weakly-supervised setups. This dataset contains a total of 141,564 expressions based on images from the COCO training set. To assess the effectiveness of CIB in addressing human-adversarial attacks, we evaluate our method on AdVQA (Sheng et al., 2021), a human-adversarial benchmark built upon VQA v2 images, featuring approximately 10k/36.8k image-question pairs for the validation/test split.

4.1.2 Evaluation Metrics

We follow previous work (Antol et al., 2015) to evaluate the VQA performance of our methods with VQA-Score. In addition, we evaluate robustness against linguistic variations using Consensus Score (CS) (Shah et al., 2019), which is the ratio of the number of subsets where all questions are answered correctly to the total number of subsets of size m. Specifically, for each question group Q containing one original question and its n corresponding rephrasings, all subsets of size m amount to ${}^{n}C_{m}$, CS can then be defined as

$$\mathrm{CS}(m) = \sum\nolimits_{q \in Q', Q' \subset Q, |Q'|=m} \frac{\mathbb{1}_{Q'}(q)}{{}^{n}C_{m}}, \tag{12}$$

where $\mathbb{1}$ is an indicator function defined on Q' and $\mathbb{1}_{Q'}(q)$ means a set where the answer to question q is correct. Naturally, the higher the average CS at larger values of m, the more robust the model. To evaluate robustness against visual variations, we utilize #flips (Agarwal et al., 2020) as a robustness evaluation metric. #flips represents the ratio of the number of prediction mismatches before and after visual content manipulation to the total number of all samples. In IV-VQA, if the predicted answers for the original image and the corresponding edited image differ, the prediction is deemed "flipped". In CV-VQA, an answer to a question based on an edited image is considered to be "flipped" if it is not one less than the prediction on the original image.

4.1.3 Baseline Pretrained VLMs

As summarized in Table 2, we utilize nine pretrained VLMs with three typical transformer architectures as baselines to evaluate the input robustness of our method. Specifically, VisualBERT (Li et al., 2019), VL-BERT$_{\mathrm{B}}$ (Su et al., 2020), and UNITER$_{\mathrm{B}}$ (Chen et al., 2020) employ single-stream encoders. LXMERT (Tan & Bansal, 2019), ViLBERT (Lu et al., 2019), and BEiT-3$_{\mathrm{B}}$ (Wang et al., 2023) utilize two-stream encoders. VL-T5 (Cho et al., 2021), ALBEF (Li et al., 2021), and mPLUG$_{\mathrm{B}}$ (Li et al., 2022) incorporate encoder-decoder architectures. When applied to the downstream VQA task, mPLUG$_{\mathrm{B}}$, VL-T5, and ALBEF formulate VQA as a text generation task (TG), while the remaining baselines formulate VQA as a multi-answer classification problem (AC). These baselines adopt two typical image tokens, namely, the region feature extracted by a pretrained object detector and the patch embedding obtained using a linear projection, and are pretrained on large-scale image-text (IT) data to learn task-agnostic versatile representations. The pretraining IT datasets include MS COCO caption (COCO) (Chen et al., 2015), Visual Genome (VG) (Krishna et al., 2017), VQA v2 (VQA) (Goyal et al., 2017), GQA balance version (GQA) (Hudson & Manning, 2019), VG-QA (VGQA) (Zhu et al., 2016), Conceptual Captions (CC) (Sharma et al., 2018), SBU captions (SBU) (Ordonez et al., 2011), and Conceptual 12 M (CC12M) (Changpinyo et al., 2021). Since VQA v2 images originate from the COCO dataset, we follow the work (Chen et al., 2020) to categorize these pretrained VLMs into in-domain (ID), in-domain and out-of-domain (ID+OOD), and

Table 2 Summary of baseline pretrained VLMs (AC: Answer Classification, TG: Text Generation)

Domain	Pretrained VLMs	Architecture	Pretraining IT Datasets	Image Tokens	VQA
ID	VisualBERT (Li et al., 2019)	Single-stream	COCO	Region feature	AC
OOD	ViLBERT (Lu et al., 2019)	Two-stream	CC	Region feature	AC
	VL-BERT$_B$ (Su et al., 2020)	Single-stream	CC	Region feature	AC
ID+OOD	VL-T5 (Cho et al., 2021)	Encoder-decoder	COCO,VG,GQA,VQA,VGQA	Region feature	TG
	LXMERT (Tan & Bansal, 2019)	Two-stream	[c]COCO,VG,GQA,VQA,VGQA	Region feature	AC
	UNITER$_B$ (Chen et al., 2020)	Single-stream	[c]COCO,VG,SUB,CC	Region feature	AC
	ALBEF (Li et al., 2021)	Encoder-decoder	COCO,VG,SUB,CC	Patch embedding	TG
	mPLUG$_B$ (Li et al., 2022)	Encoder-decoder	COCO,VG,SBU,CC,CC12M	Patch embedding	TG
	BEiT-3$_B$ (Wang et al., 2023)	Two-stream	COCO,VG,SBU,CC,CC12M	Patch embedding	AC

Table 3 Configuration setups

Methods	β	K	L	peak lr	bs
VisualBERT + CIB	5×10^{-5}	100	20	2×10^{-5}	64
ViLBERT + CIB	1×10^{-4}	10–100	20	4×10^{-5}	64
VL-BERT$_B$ + CIB	1×10^{-4}	10–100	20	4×10^{-5}	64
VL-T5 + CIB	1×10^{-4}	36	20	4×10^{-5}	64
LXMERT + CIB	5×10^{-5}	36	20	2×10^{-5}	64
UNITER$_B$ + CIB	1×10^{-4}	10–100	20	4×10^{-5}	64
ALBEF + CIB	1×10^{-4}	900	30	2×10^{-5}	32
mPLUG$_B$ + CIB	1×10^{-4}	900	80	2×10^{-5}	16
BEiT-3$_B$ + CIB	1×10^{-4}	900	64	2×10^{-5}	16

out-of-domain (OOD) groups based on whether they utilize the COCO dataset during the pretraining process.

4.1.4 Implementation Details

In the subsequent experiments, we maintain the initial configurations of all pretrained VLMs. The region features (visual inputs) of VisualBERT, VL-T5, LXMERT, UNITER$_B$, ViLBERT, and VL-BERT$_B$ are extracted using BUA Faster R-CNN (Anderson et al., 2018) pretrained on VG (Krishna et al., 2017). The representation dimension d is set to 768. The configurations of the number of word tokens L (i.e., the maximum token length allowed for a question) and image tokens K are detailed in Table 3. For the only crucial hyperparameter β in Eq. (6), it is set to 1×10^{-4} in all cases except for finetuning VisualBERT and LXMERT, where β is set to 5×10^{-5}. All experiments, except those on ALBEF, BEiT-3$_B$, and mPLUG$_B$ implemented on one NVIDIA A100 40GB GPU, are conducted using PyTorch on one NVIDIA GTX2080Ti 12GB GPU. We uniformly utilize an AdamW optimizer with a linear warmup using linear decay and a warmup step of 1000. The number of finetuning epochs is 10. The configurations of batch size and peak learning rate for each pretrained VLM are shown Table 3. The best model is selected based on the VQA-Score on the mini-split of VQA v2 training set that excludes image-question pairs when evaluating input robustness.

4.2 Input Robustness Evaluation

4.2.1 Robustness Against Linguistic Variations

To evaluate the effectiveness of CIB against linguistic variations, with CIB as the training objective, we finetune pretrained VLMs on VQA v2 training split and report results on VQA-Rephrasings and VQA P2. Tables 4 and 5 show the comparisons with existing methods in terms of the VQA-Score as well as the robustness metric of CS(m).

Result on VQA-Rephrasings We first compare the proposed CIB with existing methods: CC (Shah et al., 2019), ConClaT (Kant et al., 2021), and MANGO (Li et al., 2020). Specifically, both CC and ConClaT augment training datasets online by training a question generation model to generate paraphrases of questions. To effectively leverage augmented data and enhance model robustness to linguistic variations, CC considers cycle consistency between the question and its rephrasings, while ConClaT jointly optimizes contrastive and cross-entropy losses. CC considers three baseline VQA models, i.e., BUTD (Anderson et al., 2018), Pythia (Jiang et al., 2018), and BAN (Kim et al., 2018). ConClaT uses MMT (Kant et al., 2021), a modified version of UNITER, as

Table 4 Results of robustness against linguistic variations (i.e., sentence rephrasing) on the VQA-Rephrasings dataset (Shah et al., 2019)

Methods	VQA-Score		Robustness metric			
	PER	ORI	CS(1)	CS(2)	CS(3)	CS(4)
Data augmentation						
BUTD (Anderson et al., 2018)	51.22	61.51	60.55	46.96	40.54	34.47
+ CC (Shah et al., 2019)	52.58 (+1.36)	62.44 (+0.93)	61.66 (+1.11)	50.79 (+3.83)	44.68 (+4.14)	42.55 (+8.08)
Pythia (Jiang et al., 2018)	54.20	64.08	63.43	52.03	45.94	39.49
+ CC (Shah et al., 2019)	55.65 (+1.45)	64.52 (+0.44)	64.36 (+0.93)	55.45 (+3.42)	50.92 (+4.98)	44.30 (+4.81)
BAN (Kim et al., 2018)	55.87	64.97	64.88	53.08	47.45	39.87
+ CC (Shah et al., 2019)	56.59 (+0.72)	65.87 (+0.90)	65.77 (+0.89)	56.94 (+3.86)	51.76 (+4.31)	48.18 (+8.31)
MMT (Kant et al., 2021)	–	–	67.58	60.04	55.53	52.36
ConClaT (Kant et al., 2021)	–	–	68.62 (+1.04)	61.42 (+1.38)	57.08 (+1.55)	53.99 (+1.63)
w/o Data Augmentation						
UNITER_{B} (Chen et al., 2020)	–	–	71.29	63.95	59.48	56.31
MANGO_{B} (Li et al., 2020)	–	–	72.66 (+1.37)	66.03 (+2.08)	61.92 (+2.44)	58.95 (+2.64)
VILLA_{B} (Gan et al., 2020)	–	–	72.18	65.28	60.99	57.93
MANGO_{VB} (Li et al., 2020)	–	–	72.78 (+0.60)	65.97 (+0.69)	61.70 (+0.71)	58.59 (+0.66)
VisualBERT (Li et al., 2019)†	62.03	68.46	70.44	62.84	58.41	55.06
+ CIB	63.10 (+1.07)	69.78 (+1.32)	71.85 (+1.41)	64.16 (+1.32)	59.54 (+1.13)	56.31 (+1.25)
ViLBERT (Lu et al., 2019)†	59.16	67.65	68.00	59.65	54.68	51.22
+ CIB	62.28 (+3.12)	69.15 (+1.50)	71.05 (+3.05)	63.54 (+3.89)	59.04 (+4.36)	55.89 (+4.67)
$\text{VL-BERT}_{\text{B}}$ (Su et al., 2020)†	59.89	67.14	67.95	60.11	55.34	52.99
+ CIB	60.86 (+0.97)	68.74 (+1.60)	70.52 (+2.57)	63.46 (+3.35)	58.75 (+3.41)	53.89 (+0.90)
VL-T5 (Cho et al., 2021)†	65.64	–	71.78	65.35	62.68	61.00
+ CIB	66.93 (+1.29)	–	73.65 (+1.87)	67.48 (+2.13)	64.48 (+1.80)	62.53 (+1.53)
LXMERT (Tan & Bansal, 2019)†	70.41	–	79.73	72.93	68.49	65.21
+ CIB	**72.62** (+2.21)	–	**82.01** (+2.28)	**75.46** (+2.53)	**71.05** (+2.56)	**67.71** (+2.50)
UNITER_{B} (Chen et al., 2020)†	62.68	70.05	71.45	63.72	59.01	55.66
+ CIB	64.45 (+1.77)	70.91 (+0.86)	73.18 (+1.73)	66.21 (+2.49)	61.88 (+2.87)	58.75 (+3.09)
ALBEF (Li et al., 2021)†	65.66	71.13	70.89	65.52	61.74	60.14
+ CIB	68.00 (+2.34)	72.43 (+1.30)	73.71 (+2.82)	67.50 (+1.98)	63.60 (+1.86)	61.72 (+1.58)
mPLUG_{B} (Li et al., 2022)†	65.94	71.62	71.01	67.38	62.26	60.46
+ CIB	69.02 (+3.08)	72.86 (+1.24)	73.55 (+2.54)	70.53 (+3.15)	64.73 (+2.47)	62.95 (+2.49)
BEiT-3_{B} (Wang et al., 2023)†	67.36	73.19	75.96	69.73	65.81	62.93
+ CIB	70.01 (+2.65)	**75.06** (+1.87)	78.89 (+2.93)	73.01 (+3.28)	68.99 (+3.18)	65.92 (+2.99)

The † indicates our reimplementation of baseline, i.e., finetuning pretrained VLMs with only the task-related loss. The best performances are highlighted in bold

its baseline. MANGO employs UNITER (Chen et al., 2020) and VILLA (Gan et al., 2020) as baseline models and adopts adversarial training to enhance model robustness. As shown in Table 4,[1] the results on nine pretrained VLMs consistently show that compared to baselines (i.e., finetuning pretrained VLMs with only the task-related loss for answer prediction†), using CIB as the training objective for VQA models can significantly improve their robustness to linguistic variations. This finding suggests that it is feasible to encourage models to learn more compact and robust representations from an information-theoretic perspective. In comparison with state-of-the-art methods, adapting LXMERT with CIB achieves the best performance across all metrics. This performance advantage can be attributed to the fact that LXMERT considers the VQA training objective during pretraining, which reduces the gap between upstream and downstream objectives. In addition, we observe that the data augmentation based method (CC) yields greater improvements in the metric of CS(4). However, without data augmentation, the average improvement of our method is more substantial.

[1] VL-T5 and LXMERT utilize some examples from the VQA v2 validation set in the pretraining VQA task, resulting in an unreliable VQA-Score for ORI, thus we do not report the ORI performance.

Table 5 Results of robustness against linguistic variations (i.e., sentence rephrasing, and word substitution with synonyms and antonyms) on the VQA P2 dataset (Whitehead et al., 2020)

Methods	PER	CS(2)
Data Augmentation		
StackNMN (Hu et al., 2018)	63.30	66.20
+ Q3R (Whitehead et al., 2020)	66.90 (+3.30)	72.20 (+6.00)
HybridNet (Whitehead et al., 2020)	63.30	66.60
+ Q3R (Whitehead et al., 2020)	67.00 (+4.00)	72.50 (+5.90)
XNM (Shi et al., 2019)	64.70	68.80
+ Q3R (Whitehead et al., 2020)	68.10 (+3.40)	74.40 (+5.60)
w/o Data Augmentation		
VisualBERT (Li et al., 2019)†	68.23	72.34
+ CIB	69.92 (+1.69)	73.83 (+1.49)
ViLBERT (Lu et al., 2019)†	67.18	71.39
+ CIB	69.92 (+2.74)	73.98 (+2.59)
VL-BERT$_B$ (Su et al., 2020)†	68.36	72.52
+ CIB	69.82 (+1.46)	73.88 (+1.36)
VL-T5 (Cho et al., 2021)†	71.63	77.34
+ CIB	73.47 (+1.84)	78.99 (+1.65)
LXMERT (Tan & Bansal, 2019)†	77.30	82.96
+ CIB	**78.93** (+1.63)	**85.07** (+2.11)
UNITER$_B$ (Chen et al., 2020)†	70.36	74.36
+ CIB	71.30 (+0.94)	75.91 (+1.55)
ALBEF (Li et al., 2021)†	71.36	76.00
+ CIB	72.84 (+1.48)	77.46 (+1.46)
mPLUG$_B$ (Li et al., 2022)†	71.95	76.75
+ CIB	73.11 (+1.16)	78.09 (+1.34)
BEiT-3$_B$ (Wang et al., 2023)†	73.65	78.56
+ CIB	75.22 (+1.57)	81.28 (+2.72)

The † indicates our reimplementation of baseline, i.e., finetuning pretrained VLMs with only the task-related loss. The best performances are highlighted in bold

Result on VQA P2 We next compare our method with the existing method Q3R (Whitehead et al., 2020) on VQA P2. Q3R augments training data by creating linguistic variations such as synonymous, paraphrastic, and antonymous of input questions, and regularizes the visual reasoning process between the question and its generated questions. Q3R utilizes three baseline models: StackNMN (Hu et al., 2018), HybridNet (Whitehead et al., 2020), and XNM (Shi et al., 2019). The results in Table 5 indicate that finetuning pretrained VLMs with the proposed CIB can markedly improve their robustness against question variations on VQA P2. Moreover, finetuning LXMERT with CIB also achieves the best performance on VQA P2. In addition, the data augmentation-based method (Q3R) continues to exhibit superiority in improving the input robustness of baseline VQA models.

4.2.2 Robustness Against Visual Variations

We evaluate the robustness of our method against visual variations on IV-VQA and CV-VQA. Table 6 shows the comparisons with existing methods in the metrics of VQA-Score and #flips. CL (a simple CNN+LSTM model) (Lu et al., 2015), SNMN (an attention-based method) (Hu et al., 2018), and SAAA (a compositional model) (Kazemi & Elqursh, 2017) are benchmarked by Agarwal et al. (2020). MANGO exploits adversarial training to improve the robustness of pretrained VLMs (UNITER (Chen et al., 2020) and VILLA (Gan et al., 2020)) against visual variations. The results in Table 6 show that significant improvements are achieved across all metrics and baselines on both IV-VQA and CV-VQA, suggesting the effectiveness of CIB in improving robustness against visual variations. Moreover, we observe

Table 6 Results of robustness against visual variations on IV-VQA and CV-VQA (Agarwal et al., 2020)

Methods	IV-VQA			CV-VQA		
	PER ↑	ORI ↑	#flips ↓	PER ↑	ORI ↑	#flips ↓
CL (Lu et al., 2015)	–	60.21	17.89	–	39.38	81.41
SNMN (Hu et al., 2018)	–	66.04	6.52	–	47.95	78.92
SAAA (Kazemi & Elqursh, 2017)	–	70.26	7.85	–	49.90	78.44
UNITER_{B} (Chen et al., 2020)	–	–	8.47	–	–	40.67
MANGO_{B} (Li et al., 2020)	–	–	7.32 (+1.15)	–	–	38.11 (+2.56)
VILLA_{B} (Gan et al., 2020)	–	–	7.07	–	–	38.28
MANGO_{VB} (Li et al., 2020)	–	–	7.43 (+0.36)	–	–	38.25 (−0.03)
VisualBERT (Li et al., 2019)†	46.04	81.99	26.84	30.48	76.30	30.13
+ CIB	47.81 (+1.77)	83.48 (+1.49)	23.91 (+2.93)	32.46 (+1.98)	77.09 (+0.79)	27.98 (+2.15)
ViLBERT (Lu et al., 2019)†	72.37	81.73	11.98	32.24	70.70	35.43
+ CIB	74.67 (+2.30)	83.35 (+1.62)	10.85 (+1.13)	35.33 (+3.09)	71.11 (+0.41)	34.01 (+1.42)
$\text{VL-BERT}_{\text{B}}$ (Su et al., 2020)†	72.42	82.35	12.58	33.52	71.00	34.28
+ CIB	73.66 (+1.24)	83.37 (+1.02)	11.00 (+1.58)	35.29 (+1.77)	72.70 (+1.70)	32.09 (+2.19)
VL-T5 (Cho et al., 2021)†	75.73	–	9.76	35.09	–	33.43
+ CIB	76.46 (+0.73)	–	8.55 (+1.21)	42.55 (+7.46)	–	30.42 (+3.01)
LXMERT (Tan & Bansal, 2019)†	77.83	–	12.67	38.86	–	32.80
+ CIB	78.57 (+0.74)	–	11.64 (+1.03)	40.47 (+1.61)	–	30.37 (+2.43)
UNITER_{B} (Chen et al., 2020)†	75.71	84.56	11.77	42.60	78.27	29.67
+ CIB	76.63 (+0.92)	86.05 (+1.49)	9.80 (+1.97)	46.92 (+4.32)	79.89 (+1.62)	27.99 (+1.68)
ALBEF (Li et al., 2021)†	85.63	87.87	12.00	50.00	80.00	28.71
+ CIB	87.21 (+1.58)	88.88 (+1.01)	9.16 (+2.84)	51.78 (+1.78)	81.45 (+1.45)	25.76 (+2.95)
mPLUG_{B} (Li et al., 2022)†	86.96	89.40	13.16	53.92	78.49	25.50
+ CIB	88.47 (+1.51)	90.33 (+0.93)	10.59 (+2.57)	55.20 (+1.28)	79.43 (+0.94)	24.07 (+1.43)
BEiT-3_{B} (Wang et al., 2023)†	87.96	89.38	9.08	55.17	79.99	24.08
+ CIB	**90.00** (+2.04)	90.64 (+1.26)	**5.40** (+3.68)	**57.95** (+2.78)	81.00 (+1.01)	**23.23** (+0.85)

The † indicates our reimplementation of baseline, i.e., finetuning pretrained VLMs with only the task-related loss. The best performances are highlighted in bold

Table 7 Results of robustness against multimodal shortcut learning on the VQA-CE dataset (Dancette et al., 2021)

Methods	VQA-Score	
	CE	Easy
Shortcuts (Dancette et al., 2021)	0.00	61.13
SAN (Yang et al., 2016)	26.64	68.45
BLOCK (Ben-Younes et al., 2019)	32.91	77.65
VilBERT (Lu et al., 2019)	39.24	80.50
BUTD (Anderson et al., 2018)	33.91	76.69
+ RUBi (Cadene et al., 2019)	32.25 (−1.66)	75.03 (−1.66)
+ LMH + RMFE (Gat et al., 2020)	33.14 (−0.77)	73.32 (−3.37)
+ ESR (Shrestha et al., 2020)	33.26 (−0.65)	76.18 (−0.51)
+ LMH (Clark et al., 2019)	34.26 (+0.35)	73.12 (−3.57)
+ LfF (Nam et al., 2020)	34.27 (+0.36)	76.60 (−0.09)
+ LMH + CSS (Chen et al., 2020)	34.36 (+0.45)	62.08 (−14.61)
+ RandImg (Teney et al., 2020)	34.41 (+0.50)	76.21 (−0.48)
ViLBERT (Lu et al., 2019)†	38.91	80.96
+ CIB	41.24 (+2.33)	82.96 (+2.00)
VL-BERT$_B$ (Su et al., 2020)†	36.56	80.66
+ CIB	38.24 (+1.68)	82.00 (+1.34)
VisualBERT (Li et al., 2019)†	38.75	79.42
+ CIB	40.86 (+2.11)	81.25 (+1.83)
VL-T5 (Cho et al., 2021)†	45.41	86.05
+ CIB	47.60 (+2.19)	88.00 (+1.95)
LXMERT (Tan & Bansal, 2019)†	53.61	87.63
+ CIB	**57.14** (+3.53)	**89.21** (+1.68)
UNITER$_B$ (Chen et al., 2020)†	40.64	81.75
+ CIB	42.03 (+1.39)	82.48 (+0.73)
ALBEF (Li et al., 2021)†	45.39	83.88
+ CIB	47.87 (+2.48)	86.00 (+2.12)
mPLUG$_B$ (Li et al., 2022)†	45.74	84.07
+ CIB	47.73 (+1.99)	85.25 (+1.18)
BEiT-3$_B$ (Wang et al., 2023)†	47.15	84.44
+ CIB	50.38 (+3.23)	85.92 (+1.48)

The † indicates our reimplementation of baseline, i.e., finetuning pretrained VLMs with only the task-related loss. The best performances are highlighted in bold

that pretrained VLMs using raw images as visual inputs (e.g., BEiT-3$_B$, mPLUG$_B$, and ALBEF) exhibit superior performance in defending against visual variations compared to those pretrained VLMs (e.g., VisualBERT, LXMERT, and UNITER$_B$) that employ object-level region features as visual inputs. This can be attributed to the fact that pre-extracted region features lose some image information, which hinders VQA models in comprehending and retrieving visual content according to a given question.

4.2.3 Robustness Against Multimodal Shortcut Learning

To demonstrate the ability of CIB to defend against multimodal shortcuts present in input images and questions, we conduct experiments on VQA-CE and compare our methods with existing approaches. Results are summarized in Table 7. The compared methods in the table can be broadly classified into two groups: (*i*) plain models (SAN (Yang et al., 2016), BLOCK (Ben-Younes et al., 2019), VilBERT (Lu et al., 2019), and BUTD (Anderson et al., 2018)), and (*ii*) bias-reduction methods (RUBi (Cadene et al., 2019), LMH + RMFE (Gat et al., 2020), ESR (Shrestha et al., 2020), LMH (Clark et al., 2019), LfF (Nam et al., 2020), LMH + CSS (Chen et al., 2020), and RandImg (Teney et al., 2020)). These experimental results are cited from the work (Dancette et al., 2021). As shown in Table 7, finetuning baseline pretrained VLMs with CIB achieves significant improvements and outperforms bias-reduction methods by a considerable margin, particularly on counterexamples. These results suggest that the proposed CIB is more effective at alleviating the spurious

Table 8 Comparison between different CIB bounds

VLMs	CIB Terms					PER
	$I(Y;T)$	$I(X^v;T^v)$	$I(X^l;T^l)$	$-I(T^v;T^l)$	D_{skl}	
LXMERT	✓					70.41
	✓	✓	✓			72.17
	✓			✓	✓	72.07
	✓	✓	✓		✓	72.28
	✓	✓	✓	✓	✓	**72.62**
UNITER_{B}	✓					62.68
	✓	✓	✓			64.07
	✓			✓	✓	64.11
	✓	✓	✓		✓	63.23
	✓	✓	✓	✓	✓	**64.45**
ALBEF	✓					65.66
	✓	✓	✓			67.11
	✓			✓	✓	67.00
	✓	✓	✓		✓	66.84
	✓	✓	✓	✓	✓	**68.00**

The best results for each baseline are highlighted in bold

correlations between representations and reducing shortcut learning involved in multimodal inputs.

4.3 Ablation Studies

4.3.1 Comparison with Alternative CIB Bounds

When finetuning pretrained VLMs with CIB, $I(Y;T)$ is regarded as the task-related objective, while $I(X^v, X^l; T^v, T^l)$ serves as a MI regularizer to constrain representation compactness and pursue more robust representations. As stated in Theorem 1, the upper bound of $I(X^v, X^l; T^v, T^l)$ consists of four terms: $I(X^v;T^v)$, $I(X^l;T^l)$, $-I(T^v;T^l)$, and D_{skl}. To analyze the contribution of different terms to CIB, we perform an ablation study on different meaningful combinations of these terms, that is, provable upper bounds of $I(X^v, X^l; T^v, T^l)$, on VQA-Rephrasings using LXMERT, UNITER_{B}, and ALBEF as baseline pretrained VLMs. Specifically, the regularizer upper bound has three other meaningful alternatives: (*i*) $\frac{3}{2}[I(X^v;T^v)+I(X^l;T^l)]$, (*ii*) $-I(T^v;T^l)+D_{\text{skl}}$, and (*iii*) $I(X^v;T^v)+I(X^l;T^l)+D_{\text{skl}}$ (cf. Sect. 1 for proofs). Table 8 presents results on VQA-Rephrasings. Overall, the ablation results on different bounds are consistent, indicating that CIB with any meaningful upper bounds can markedly improve the performance of baseline pretrained VLMs. However, CIB with our derived upper bound performs best, empirically demonstrating that the bound in Theorem 1 is a tighter and more precise bound. Furthermore, the comparison between upper bound (*iii*) $I(X^v;T^v)+I(X^l;T^l)+D_{\text{skl}}$ and our upper bound $I(X^v;T^v)+I(X^l;T^l)-I(T^v;T^l)+D_{\text{skl}}$ suggests that CIB can effectively facilitate the correlation between visual and linguistic representations and modality alignment by maximizing $I(T^v;T^l)$.

4.3.2 Impact of MI Estimator on CIB

In practice, any sample-based upper bound estimator of MI can be utilized to approximate $I(X^v;T^v)$ and $I(X^l;T^l)$, and any differentiable MI lower bound estimator can be applied to approach $I(T^v;T^l)$. To analyze the impact of different MI estimators on CIB, we consider the following experimental settings: (*i*) We alternately utilize L1Out (Poole et al., 2019) instead of CLUB (Cheng et al., 2020) as the estimator of MI upper bound to approximate $I(X^v;T^v)$ and $I(X^l;T^l)$. (*ii*) We approximate $I(T^v;T^l)$ with the three other estimators of MI lower bound, i.e., InfoNCE (Oord et al., 2018), NWJ (Nguyen et al., 2010), and MINE (Belghazi et al., 2018). Table 9 presents comparisons between different MI estimators on VQA-Rephrasings using LXMERT, UNITER_{B}, and ALBEF as baselines. These results consistently demonstrate that CIB can effectively improve the performance of baselines with different transformer architectures and that the effectiveness of CIB does not depend on a specific MI estimator.

4.3.3 Impact of Hyperparameter on CIB

When using CIB as the training objective to adapt pretrained VLMs to the downstream VQA task, β controls the tradeoff between redundancy and compression in representations, which is the crucial hyperparameter. Consequently, we perform a grid search for β. Specifically, we consider the following values: $\beta \in [1\times10^{-6}, 1\times10^{-5}, 5\times10^{-5}, 1\times$

Table 9 Impact of different MI estimators on CIB

VLMs	MI Estimator		*PER
	Upper Bound	Lowe Bound	
LXMERT			70.41
	L1Out	NWJ	72.31
	CLUB	InfoNCE	72.34
	CLUB	MINE	72.48
	CLUB	NWJ	**72.62**
UNITER_{B}			62.68
	L1Out	NWJ	64.27
	CLUB	InfoNCE	64.14
	CLUB	MINE	64.32
	CLUB	NWJ	**64.45**
ALBEF			65.66
	L1Out	NWJ	67.68
	CLUB	InfoNCE	**68.00**
	CLUB	MINE	67.91
	CLUB	NWJ	**68.00**

The best results for each baseline are highlighted in bold

$10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-3}, 5 \times 10^{-3}, 1 \times 10^{-2}, 5 \times 10^{-2}]$. Figure 3 illustrates the variation curve of VQA-Score (PER) on VQA-Rephrasings with increasing $\log \beta$. We observe that the performance starts to boost when β is quite small, indicating the effectiveness of CIB in improving the performance of baseline pretrained VLMs. When β increases to 5×10^{-5}, 1×10^{-4}, and 1×10^{-4}, UNITER_{B}, LXMERT, and ALBEF respectively achieve the best performance. Beyond that point, the performance typically begins to degrade, suggesting that extremely compressed representations of pretrained VLMs may start to compromise model performance.

4.3.4 Impact of Internal Representation on CIB

As illustrated in Fig. 1b, for pretrained VLMs with two-stream encoders (e.g., LXMERT and ViLBERT), there is an alternative option for internal representations, i.e., $T = [T^{'v}, T^{'l}]$, which are the visual and linguistic representations after the vision transformer layers ($f_{\theta^{\text{VTran}}}$) and language transformer layers ($f_{\theta^{\text{LTran}}}$). When finetuning the two-stream pretrained VLMs with CIB, we analyze the impact of different internal representations by replacing the original $T = [T^{v}, T^{l}]$ in $\mathscr{L}_{\text{CIB}}$ with $T = [T^{'v}, T^{'l}]$. Table 10 shows the VQA-Score for PER on VQA-Rephrasings, revealing that different internal representations have a slight impact on the PER performance of CIB. This indicates that for two-stream pretrained VLMs, using the visual and linguistic representations after the vision and language transformer layers as internal representations to estimate the mutual information terms in $\mathscr{L}_{\text{CIB}}$ is also a feasible approach.

4.4 Discussion and Analysis

4.4.1 Effectiveness of CIB for Standard VQA Performance

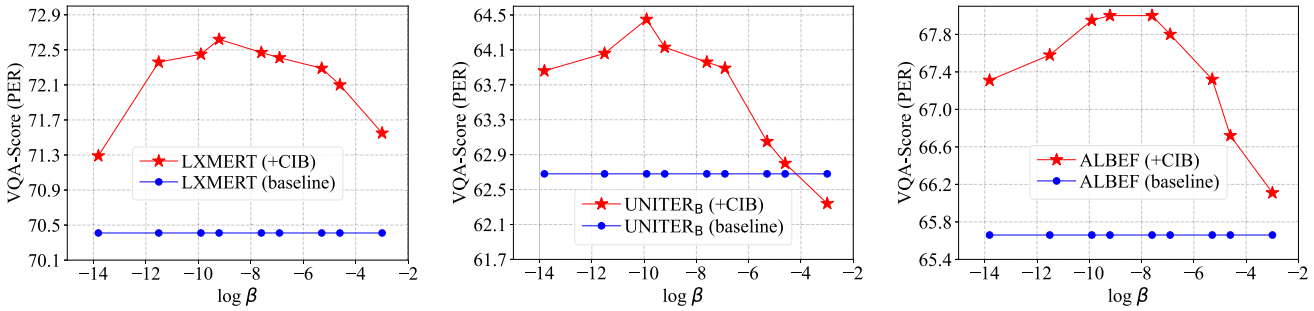
To analyze the impact of CIB on standard VQA performance (i.e., whether the representation compression impairs the standard VQA performance), we utilize CIB as the objective to train the aforementioned baseline pretrained VLMs on the VQA v2 training and validation sets. The results on VQA v2 test-dev are shown in Table 11.[2] Overall, training baseline pretrained VLMs with the proposed CIB can slightly improve the standard VQA performance. In particular, the performance improvement of VisualBERT and ALBEF is relatively significant. This because that their visual inputs contain more redundant information, such as image background and visual content irrelevant to the given question (VisualBERT and ALBEF respectively adopt 100 region-level features and 900 patch-level features as visual inputs). Therefore, our hypothesis is that a certain degree of compression of representations can reduce the redundant information learned from inputs and make the obtained representations more compact and robust. Noting that, in contrast to the significant improvement in input robustness, CIB leads to relatively limited improvement in standard performance. This observation also indirectly indicates that the proposed CIB is carefully designed and tailored to improve input robustness when adapting pretrained VLMs for the downstream VQA.

4.4.2 Effectiveness of CIB Against Adversarial Attack

To analyze the impact of CIB on defending against adversarial attacks, we conduct experiments considering the following attacks and dataset: (*i*) L4A (Ban & Dong, 2022), which adds pretrained adversarial perturbations (PAPs) to the low-level layer of pretrained models, can effectively fool the finetuned models on downstream tasks without any knowledge of the tasks. (*ii*) AdVQA (Sheng et al., 2021), an adversarial benchmark collected using a human-and-model-in-the-loop paradigm to attack state-of-the-art VQA models and obtain human-adversarial examples, can effectively evaluate the human-adversarial robustness of VQA models. The effectiveness of a model in defending against adversarial attacks is measured by the model VQA-Score under these attacks.

For (*i*), as proposed in the work (Ban & Dong, 2022), we consider three different ways for perturbation generation, i.e., L4A_{base}, L4A_{fuse}, and L4A_{ugs}. Before finetuning the pre-

[2] Since we do not utilize the additional question-answer pairs from VG (Krishna et al., 2017) for data augmentation in our experiments and some other detail differences, there are minor differences between our re-implementation[†] of the baseline pretrained VLMs and the published results in the original papers.



(a) LXMERT (b) UNITER$_B$ (c) ALBEF

Fig. 3 Variation curve of VQA-Score (PER) on VQA-Rephrasings as log β increases

Table 10 Impact of different internal representations obtained by the two-stream pretrained VLMs on CIB

VLMs	$\mathscr{L}_{\text{CIB}}$	PER
LXMERT	$I(Y;T)$	70.41
	$I(Y;T)-\beta I(X^v,X^l;T'^v,T'^l)$	72.53
	$I(Y;T)-\beta I(X^v,X^l;T^v,T^l)$	**72.62**
ViLBERT	$I(Y;T)$	59.16
	$I(Y;T)-\beta I(X^v,X^l;T'^v,T'^l)$	62.23
	$I(Y;T)-\beta I(X^v,X^l;T^v,T^l)$	**62.28**

The best results for each baseline are highlighted in bold

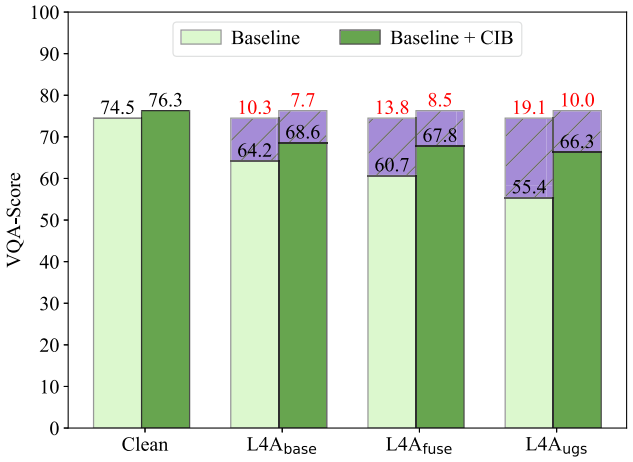


Fig. 4 Results on VQA v2 test-dev under different adversarial attacks

trained ALBEF with a text generation loss and the proposed CIB as a training objective, we first utilize the three methods above to generate PAPs by lifting the neuron activations of low-level layers of the ALBEF. Next, we separately add the generated PAPs to input images and finetune the pretrained ALBEF on the VQA v2 training and validation sets, and test their performance on VQA v2 test-dev. Figure 4 shows the performance comparison, where the blue bar marked with red performance indicates the VQA-Score drop with respect to the standard performance under an attack. From the figure, we can observe that CIB markedly reduces the performance drop, demonstrating its ability to better alleviate the vulnerability of VQA models to such attacks. For (*ii*), following the experimental setups for evaluating input robustness, we first finetune pretrained UNITER$_B$ and LXMERT on the standard and clean VQA v2 training set, and then evaluate human-adversarial robustness on the adversarial benchmark AdVQA. As shown in Table 12, the significant performance improvement of our method over baselines demonstrates the robustness of CIB against human-adversarial attacks. In summary, the aforementioned experiments consistently suggest that the proposed CIB, as a generic objective, can potentially

Table 11 Results on VQA v2 test-dev (Goyal et al., 2017) under standard and clean dataset setups

VLMs	VQA-Score	
	Baseline	+ CIB
VisualBERT (Li et al., 2019)	70.80 (70.46†)	71.62 (+1.16)
VL-T5 (Cho et al., 2021)	- (70.23†)	71.14 (+0.91)
LXMERT (Tan & Bansal, 2019)	72.42 (72.58†)	72.99 (+0.41)
UNITER$_B$ (Chen et al., 2020)	72.70 (71.63†)	72.11 (+0.48)
ALBEF (Li et al., 2021)	74.54 (74.54†)	76.27 (+1.73)
ViLBERT (Lu et al., 2019)	70.55 (70.55†)	71.00 (+0.45)
VL-BERT$_B$ (Su et al., 2020)	71.16 (71.20†)	71.59 (+0.39)

The † indicates our reimplementation of baseline, i.e., finetuning pretrained VLMs with only the task-related loss

Table 12 Results on AdVQA (Sheng et al., 2021)

Methods	VQA-Score	
	Test	Val
VisualBERT (Li et al., 2019)	31.96	28.09
ViLBERT (Lu et al., 2019)	32.01	33.67
ViLT (Kim et al., 2021)	31.00	32.48
UNITER_{B} (Chen et al., 2020)	27.56	29.44
VILLA_{B} (Gan et al., 2020)	27.55	29.36
UNITER_{L} (Chen et al., 2020)	29.66	32.08
VILLA_{L} (Gan et al., 2020)	28.59	30.58
M4C (Hu et al., 2020)	36.57	36.93
UNITER_{B} (Chen et al., 2020)†	36.20	36.73
+ CIB	37.85 (+1.65)	38.23 (+1.50)
LXMERT (Tan & Bansal, 2019)†	36.30	37.09
+ CIB	**37.42** (+1.12)	**39.10** (+2.01)

The † indicates our reimplementation of baseline, i.e., finetuning pretrained VLMs with only the task-related loss. The best performances are highlighted in bold

Table 13 Results on the RefCOCO+ (Yu et al., 2016) dataset for weakly-supervised visual grounding

Methods	Val	Test A	Test B
ARN (Liu et al., 2019)	32.78	34.35	32.13
CCL (Zhang et al., 2020)	34.29	36.91	33.56
$\text{ALBEF}_{\text{itc}}$ (Li et al., 2021)	51.58	60.09	40.19
$\text{ALBEF}_{\text{itm}}$ (Li et al., 2021)	58.46	65.89	46.25
+ CIB	**59.41**	**67.39**	**47.18**

The best performances are highlighted in bold

alleviate the vulnerability of models to adversarial attacks when adapting pretrained VLMs to downstream tasks.

4.4.3 Generalizability of CIB to Other Multimodal Task

The proposed CIB is essentially a generic training objective that can be applied to various multimodal tasks beyond the VQA task. To evaluate the generalizability of CIB to other multimodal tasks, we consider the task of weakly-supervised visual grounding. Following the original experimental setups of ALBEF (Li et al., 2021), we finetune pretrained ALBEF with CIB on the RefCOCO+ (Yu et al., 2016) training dataset in a weakly-supervised setups, i.e., finetuning models using only image-text supervision without bounding box annotations. From the results in Table 13, we find that using CIB as the training objective can further improve the performance of the baseline $\text{ALBEF}_{\text{itm}}$, demonstrating that the proposed CIB can be effectively applied to other multimodal tasks.

4.5 Qualitative Results

4.5.1 Visualization of Visual Attentional Objects

To empirically explore why CIB can improve input robustness, we utilize the pretrained LXMERT (Tan & Bansal, 2019) as a representative and conduct the following experiments. First, we enumerate the image-question pairs, whose answers are correctly predicted by the LXMERT finetuned with CIB but incorrectly predicted by the baseline LXMERT finetuned without CIB, from the VQA-Rephrasings dataset. Next, we compute the attention score between the final representation $Z \in \mathbb{R}^d$ used for answer prediction and the input visual representation $X^v \in \mathbb{R}^{K\times d}$ of object regions using the formula, i.e., $\text{score}_{\text{attn}} = \text{softmax}(Z \cdot (X^v)^{\mathbb{T}}/\sqrt{d})$. Finally, we utilize the magenta and green color to highlight the top two objects with the highest attention scores in the

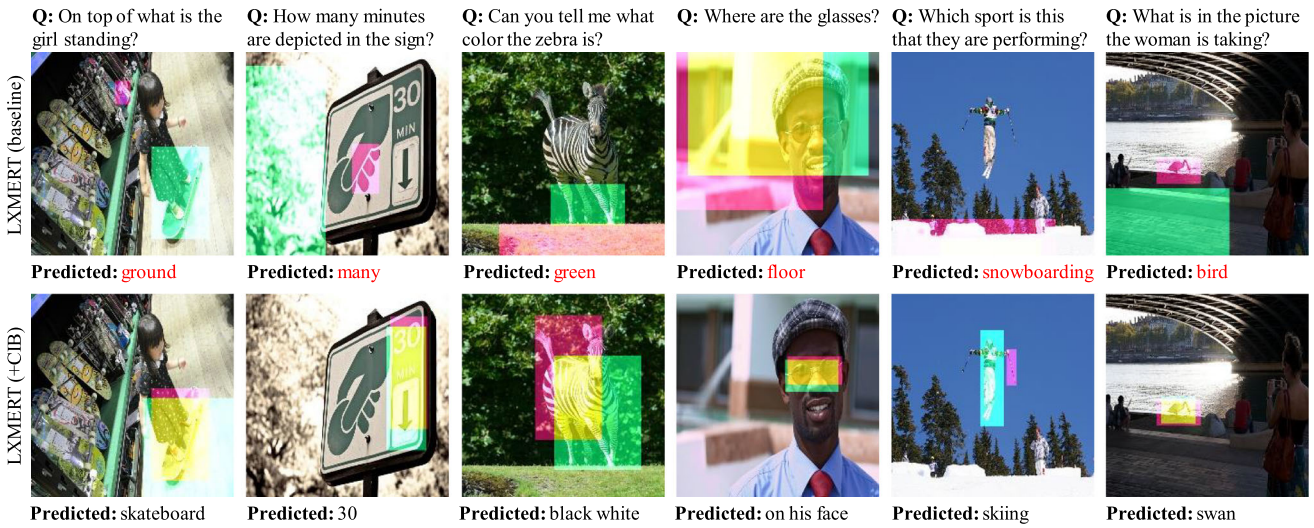
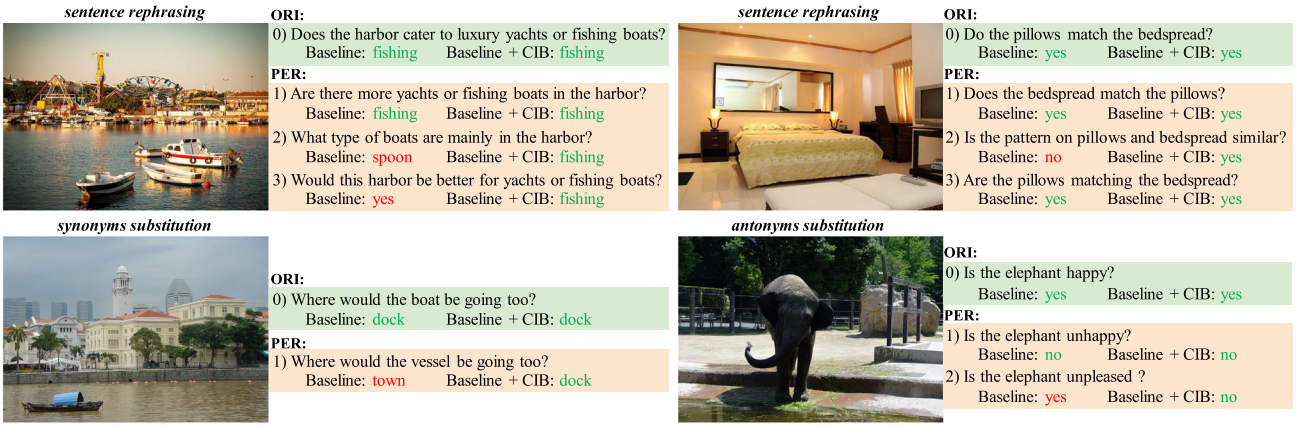
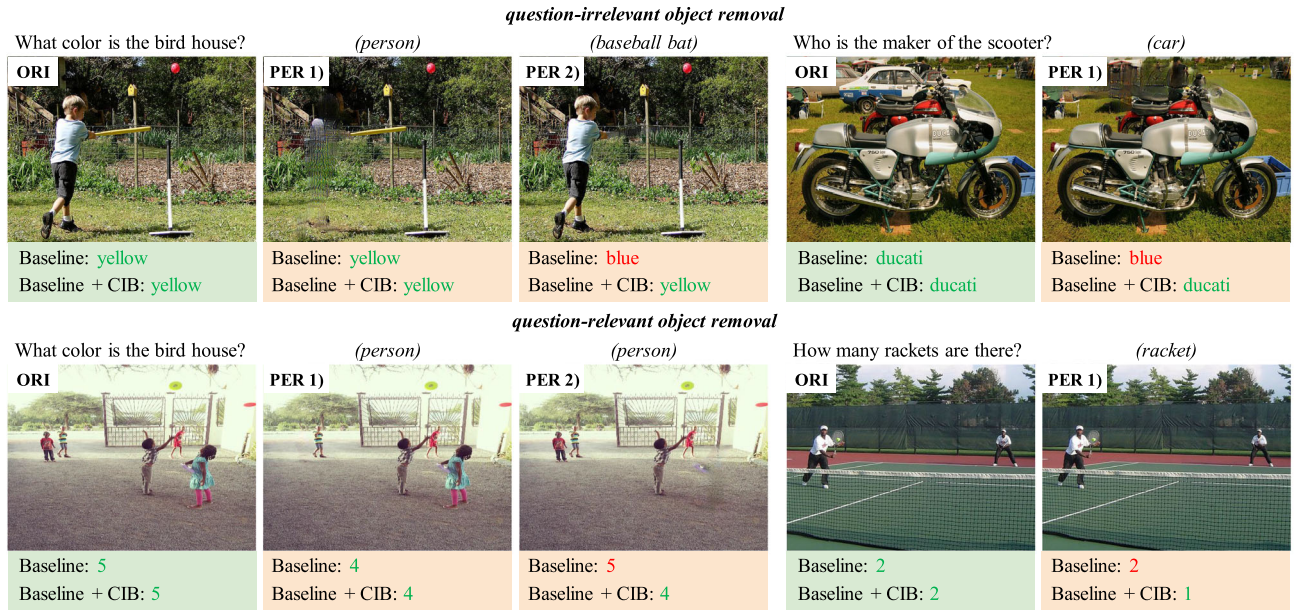


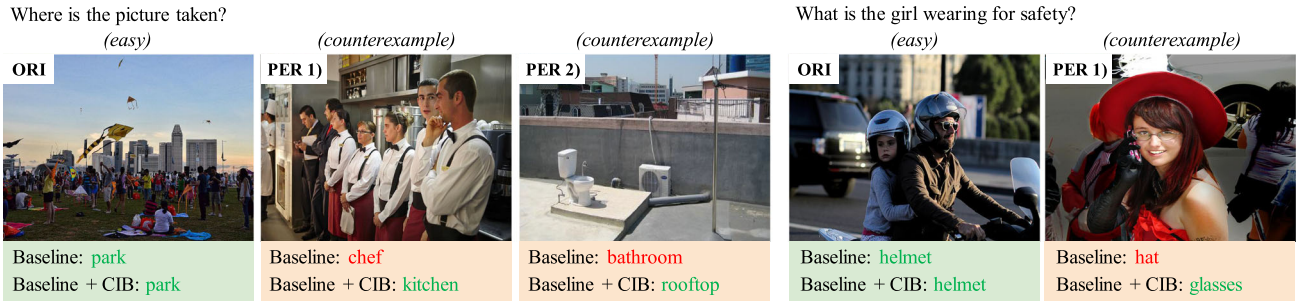
Fig. 5 Visualization of the top two objects with the highest attention scores. The image-question pairs originate from VQA-Rephrasings. Objects with the best and second attention scores are marked in magenta and green. The wrongly predicted answers are marked in red (Color figure online)



(a) Examples of robustness against linguistic variants on VQA-Rephrasings and VQA P2



(b) Examples of robustness against viaual variants on IV-VQA and CV-VQA



(c) Examples of robustness against multimodal shotcut learning on VQA-CE

Fig. 6 Qualitative examples of the baseline and our method. The correct and wrong answers are highlighted (Color figure online)

image. The results in Fig. 5 show that compared with the baseline LXMERT, the attended two objects obtained by the LXMERT finetuned with CIB are more consistent and question-related. This observation qualitatively illustrates that using CIB as a training objective to finetune pretrained VLMs can encourage models to learn more discriminative representations for different answers and reduce the irrelevant information to questions.

4.5.2 Visualization of Input Robustness Cases

Figure 6a, b, and c present several qualitative examples demonstrating robustness to linguistic variations, visual variations, and multimodal shortcut learning, respectively. According to the qualitative comparison in Fig. 6, we can further observe that compared to the baseline that finetunes the pretrained LXMERT with a cross-entropy loss, using the proposed CIB as a training objective to finetune the pretrained VLM can improve the ability of VQA models to correctly answer these difficult questions. This empirical evidence highlights the effectiveness of CIB in defending against such attacks involving both visual and linguistic inputs.

5 Conclusion

In this paper, we propose to improve input robustness from the information bottleneck perspective when adapting pretrained VLMs to the downstream VQA task. Specifically, we derive a new IB lower bound (CIB) for vision-language learning and apply CIB to finetune pretrained VLMs with various architectures for VQA. Extensive experiments on five robustness datasets consistently demonstrate the effectiveness and superiority of CIB. In the future, we plan to assess the effectiveness of CIB when tuning pretrained VLMs using parameter-effective strategies, such as adapter-based tuning and prompt-based tuning.

Limitation. Redundancy has two sides. One reason why pretrained VLMs can significantly improve the performance of downstream tasks is that they have learned rich and redundant knowledge during the pretraining stage. Practically, for downstream tasks, especially in-domain tasks, task-related redundancy can help models quickly adapt to new tasks, while task-agnostic redundancy may impair model robustness. Our work investigates improving input robustness of models while preserving their accuracy by seeking a trade-off between representation compression and redundancy. Another potential research direction is to explore how to explicitly reduce task-agnostic redundancy and adequately exploit task-related redundancy when adapting pretrained VLMs to downstream tasks, particularly out-of-domain tasks.

Acknowledgements This work was supported by the National Science Foundation of China (Grant No. 62088102).

A Appendix

A.1 Proof for Theorem 1

To prove the Theorem 1 stated in Sect. 3.2, we first enumerate some properties of mutual information (MI). Specifically, for any random variables X, Y and Z, we have:

(P_1) Positivity:

$$I(X;Y) \geq 0, I(X;Y|Z) \geq 0.$$

(P_2) Chain rule:

$$\begin{aligned} &I(X,Y;Z) \\ &= I(Y;Z) + I(X;Z|Y), \\ &= I(X;Z) + I(Y;Z|X), \\ &= \frac{1}{2}\left[I(Y;Z) + I(X;Z) + I(X;Z|Y) + I(Y;Z|X)\right]. \end{aligned}$$

(P_3) Chain rule (Multivariate Mutual Information):

$$I(X;Y;Z) = I(Y;Z) - I(Y;Z|X).$$

(P_4) Positivity of discrete entropy (for discrete X):

$$H(X) \geq 0, H(X|Y) \geq 0.$$

(P_5) Entropy and Mutual Information:

$$H(X) = H(X|Y) + I(X;Y).$$

Then, we state the following three easily provable lemmas:

Lemma 1 *In representation learning, given a random variable X, the random variable Z is defined to be a representation of X, we can simply state that Z is conditionally independent from any other variable in the model once X is observed. That is, for any variable (or groups of variables) T_1 and T_2 in the model, we have*

$$I(Z;T_1|X,T_2) = 0.$$

Lemma 2 *Given a sequence of random variables X_1, X_2, $\ldots$, X_n and a deterministic function f, then $\forall i, j = 1, 2, \ldots, n$, we have*

$$I(X_i; f(X_i)) \geq I(X_j; f(X_i)).$$

Proof By the definition,

$$\begin{aligned} I(X_i; f(X_i)) &= H(f(X_i)) - H(f(X_i) \mid X_i), \\ I(X_j; f(X_i)) &= H(f(X_i)) - H(f(X_i) \mid X_j). \end{aligned}$$

Since f is a deterministic function,

$$H(f(X_i) \mid X_i) = 0,$$

$H(f(X_i) \mid X_j) \geq 0.$

Therefore,

$I(X_i; f(X_i)) \geq I(X_j; f(X_i)).$

□

Lemma 3 *Let Z_1 and Z_2 are the representations of X_1 and X_2, then*

$$I_\theta(X_1; Z_1|X_2) \leq KL\left(p_\theta(Z_1|X_1)||p_\psi(Z_2|X_2)\right),$$
$$I_\psi(X_2; Z_2|X_1) \leq KL\left(p_\psi(Z_2|X_2)||p_\theta(Z_1|X_1)\right).$$

Proof By the definition,

$$\begin{aligned}
&I_\theta(X_1; Z_1|X_2)\\
&= \mathbb{E}_{x_1,x_2\sim p(X_1,X_2)}\mathbb{E}_{z\sim p_\theta(Z_1|X_1)}\left[\log\frac{p_\theta(Z_1=z|X_1=x_1)}{p_\theta(Z_1=z|X_2=x_2)}\right],\\
&= \mathbb{E}_{x_1,x_2\sim p(X_1,X_2)}\mathbb{E}_{z\sim p_\theta(Z_1|X_1)}\left[\log\frac{p_\theta(Z_1=z|X_1=x_1)}{p_\psi(Z_2=z|X_2=x_2)}\right]\\
&- \mathbb{E}_{x_1,x_2\sim p(X_1,X_2)}\mathbb{E}_{z\sim p_\theta(Z_1|X_1)}\left[\log\frac{p_\theta(Z_1=z|X_2=x_2)}{p_\psi(Z_2=z|X_2=x_2)}\right],\\
&= \mathrm{KL}\left(p_\theta(Z_1|X_1)||p_\psi(Z_2|X_2)\right)\\
&\quad - \mathrm{KL}\left(p_\theta(Z_2|X_1)||p_\psi(Z_2|X_2)\right),\\
&\leq \mathrm{KL}\left(p_\theta(Z_1|X_1)||p_\psi(Z_2|X_2)\right).
\end{aligned}$$

If and only if $p_\psi(Z_2|X_2)$ coincides with $p_\theta(Z_1|X_2)$, the equality holds. Analogously, $I_\psi(X_2; Z_2|X_1) \leq \mathrm{KL}(p_\psi(Z_2|X_2)||p_\theta(Z_1|X_1))$ is proved. □

Next, we utilize the above properties and lemmas to prove Theorem 1.

Theorem 1 *(Upper Bound of $I(X^v, X^l; T^v, T^l)$) Given two groups of random variables $X = [X^v, X^l]$ and $T = [T^v, T^l]$, $I(X^v, X^l; T^v, T^l)$ can be upper-bounded with*

$$\begin{aligned}
I(X; T) &= I(X^v, X^l; T^v, T^l),\\
&\leq I(X^v; T^v) + I(X^l; T^l) - I(T^v; T^l) + D_{\mathrm{skl}},
\end{aligned}$$

where D_{skl} denotes the symmetric Kullback-Leibler (KL) divergence and can be obtained by averaging the divergences $\mathrm{KL}(p(t^v|x^v)||p(t^l|x^l))$ and $\mathrm{KL}(p(t^l|x^l)||p(t^v|x^v))$.

Proof

$$\begin{aligned}
&I(X; T)\\
&= I(X^l, X^v; T),\\
&\overset{(P_2)}{=} \frac{1}{2}\left[I(X^l; T) + I(X^v; T) + I(X^l; T|X^v) + I(X^v; T|X^l)\right],\\
&= \frac{1}{2}\left[I(X^l; T^l, T^v) + I(X^v; T^l, T^v) + I(X^l; T|X^v) + I(X^v; T|X^l)\right].
\end{aligned}$$

Since,

$$\begin{aligned}
&I(X^l; T^l, T^v)\\
&\overset{(P_2)}{=} I(X^l; T^l) + I(X^l; T^v|T^l),\\
&\overset{(P_3)}{=} I(X^l; T^l) + I(X^l; T^v) - I(X^l; T^v; T^l),\\
&\overset{(P_3)}{=} I(X^l; T^l) + I(X^l; T^v) - I(T^l; T^v) + I(T^l; T^v|X^l),\\
&\overset{(LA1)}{=} I(X^l; T^l) + I(X^l; T^v) - I(T^l; T^v),\\
&\overset{(LA2)}{\leq} 2I(X^l; T^l) - I(T^l; T^v).
\end{aligned}$$

Analogously, $I(X^v; T^l, T^v)$ is upper bounded by

$$I(X^v; T^l, T^v) \leq 2I(X^v; T^v) - I(T^l; T^v).$$

And,

$$\begin{aligned}
&I(X^l; T|X^v)\\
&= I(X^l; T^l, T^v|X^v),\\
&\overset{(P_2)}{=} I(X^l; T^l|X^v) + I(X^l; T^v|X^v, T^l),\\
&\overset{(LA1)}{=} I(X^l; T^l|X^v);\\
&I(X^v; T|X^l)\\
&= I(X^v; T^l, T^v|X^l),\\
&\overset{(P_2)}{=} I(X^v; T^v|X^l) + I(X^v; T^l|X^l, T^v),\\
&\overset{(LA1)}{=} I(X^v; T^v|X^l).
\end{aligned}$$

Let $D_{\mathrm{skl}} = \frac{1}{2}\left(\mathrm{KL}(p_\theta||p_\psi) + \mathrm{KL}(p_\psi||p_\theta)\right)$, therefore,

$$\begin{aligned}
&I(X; T)\\
&= I(X^l, X^v; T^l, T^v),\\
&\leq I(X^l; T^l) + I(X^v; T^v) - I(T^l; T^v)\\
&\quad + \frac{1}{2}\left[I(X^l; T^l|X^v) + I(X^v; T^v|X^l)\right],\\
&\overset{(LA3)}{\leq} I(X^l; T^l) + I(X^v; T^v) - I(T^l; T^v)\\
&\quad + \frac{1}{2}\left[\mathrm{KL}\left(p_\theta(T_1|X_1)||p_\psi(T_2|X_2)\right)\right.\\
&\quad \left.+\mathrm{KL}\left(p_\psi(T_2|X_2)||p_\theta(T_1|X_1)\right)\right],\\
&= I(X^l; T^l) + I(X^v; T^v) - I(T^l; T^v) + D_{\mathrm{skl}}.
\end{aligned}$$

□

A.2 Proof for Alternative Upper Bound

The alternative three upper bounds of $I(X^v, X^l; T^v, T^l)$ utilized in Sect. 4.3.1 are derived as follows:

Theorem 3 *(Upper Bound of $I(X^v, X^l; T^v, T^l)$) Given two groups of random variables $X = [X^v, X^l]$ and $T = [T^v, T^l]$, the mutual information $I(X^v, X^l; T^v, T^l)$ can be upper-bounded with*

$$I(X^v, X^l; T^v, T^l) \leq \frac{3}{2}\Big[I(X^v; T^v) + I(X^l; T^l)\Big].$$

Proof

$$\begin{aligned}
&I(X^l, X^v; T^l, T^v)\\
&\leq I(X^l; T^l) + I(X^v; T^v) - I(T^l; T^v)\\
&\quad + \frac{1}{2}\Big[I(X^l; T^l|X^v) + I(X^v; T^v|X^l)\Big],\\
&\leq I(X^l; T^l) + I(X^v; T^v)\\
&\quad + \frac{1}{2}\Big[I(X^l; T^l|X^v) + I(X^v; T^v|X^l)\Big],\\
&\leq I(X^l; T^l) + I(X^v; T^v) + \frac{1}{2}\Big[I(X^l; T^l) + I(X^v; T^v)\Big],\\
&= \frac{3}{2}\Big[I(X^l; T^l) + I(X^v; T^v)\Big].
\end{aligned}$$

□

Theorem 4 *(Upper Bound of $I(X^v, X^l; T^v, T^l)$) Given two groups of random variables $X = [X^v, X^l]$ and $T = [T^v, T^l]$, the mutual information $I(X^v, X^l; T^v, T^l)$ can be upper-bounded with*

$$I(X^v, X^l; T^v, T^l) \leq I(X^v; T^v) + I(X^l; T^l) + D_{skl}.$$

where D_{skl} denotes symmetric Kullback–Leibler (KL) divergence and can be obtained by averaging the divergences $KL(p(t^v|x^v)||p(t^l|x^l))$ and $KL(p(t^l|x^l)||p(t^v|x^v))$.

Proof

$$\begin{aligned}
&I(X^l, X^v; T^l, T^v)\\
&\overset{\text{(Theorem 1)}}{\leq} I(X^l; T^l) + I(X^v; T^v) - I(T^l; T^v) + D_{\text{skl}},\\
&\overset{(I(T^l;T^v)\geq 0)}{\leq} I(X^l; T^l) + I(X^v; T^v) + D_{\text{skl}}.
\end{aligned}$$

□

Theorem 5 *(Upper Bound of $I(X^v, X^l; T^v, T^l)$) Given two groups of random variables $X = [X^v, X^l]$ and $T = [T^v, T^l]$, the mutual information $I(X^v, X^l; T^v, T^l)$ can be upper-bounded with*

$$I(X^v, X^l; T^v, T^l) \leq -I(T^v; T^l) + D_{skl}.$$

where D_{skl} denotes symmetric Kullback–Leibler (KL) divergence and can be obtained by averaging the divergences $KL(p(t^v|x^v)||p(t^l|x^l))$ and $KL(p(t^l|x^l)||p(t^v|x^v))$.

Please see the work of Federici et al. (2020) for proof.

A.3 Proof for Theoretical Justification of Input Robustness

Finally, we prove the theoretical justification for the input robustness of CIB in Eq. (11), i.e., the following inequality:

$$\begin{aligned}
&|I(T; Y) - I(T'; Y)|\\
&= |I(T^v, T^l; Y) - I(T^{v\prime}, T^{l\prime}; Y)|,\\
&\leq B_1^v\sqrt{\mathcal{T}^v}\left(I(X^v; T^v)\right)^{1/2} + B_2^v|\mathcal{T}^v|^{3/4}\left(I(X^v; T^v)\right)^{1/4}\\
&\quad + B_3^v\sqrt{|\mathcal{T}^v|}\left(I(X^{v\prime}; T^{v\prime})\right)^{1/2} + B_4^v|\mathcal{T}^v|^{3/4}\left(I(X^{v\prime}; T^{v\prime})\right)^{1/4}\\
&\quad + B_1^l\sqrt{\mathcal{T}^l}\left(I(X^l; T^l)\right)^{1/2} + B_2^l|\mathcal{T}^l|^{3/4}\left(I(X^l; T^l)\right)^{1/4}\\
&\quad + B_3^l\sqrt{|\mathcal{T}^l|}\left(I(X^{l\prime}; T^{l\prime})\right)^{1/2} + B_4^l|\mathcal{T}^l|^{3/4}\left(I(X^{l\prime}; T^{l\prime})\right)^{1/4}\\
&\quad + B_0^v + B_0^l,
\end{aligned}$$

where $\mathcal{T}^v$ is the finite support of T^v and $T^{v\prime}$, and B_0^v, B_1^v, B_2^v, B_3^v, and B_4^v are constants that depend on the sequence length K, δ, and $p(x^v)$. $\mathcal{T}^l$ is the finite support of T^l and $T^{l\prime}$, and B_0^l, B_1^l, B_2^l, B_3^l, and B_4^l are constants that depend on the sequence length L, δ, and $p(x^l)$.

Proof According to triangle inequality and data processing inequality:

$$\begin{aligned}
&|I(T; Y) - I(T'; Y)|\\
&= |I(T^v, T^l; Y) - I(T^{v\prime}, T^{l\prime}; Y)|,\\
&= |I(T^v; Y) + I(T^l; Y|T^v) - I(T^{v\prime}; Y) - I(T^{l\prime}; Y|T^{v\prime})|,\\
&= |I(T^v; Y) - I(T^{v\prime}; Y)| + |I(T^l; Y|T^v) - I(T^{l\prime}; Y|T^{v\prime})|,\\
&\leq |I(T^v; Y) - I(T^{v\prime}; Y)| + |I(T^l; Y) - I(T^{l\prime}; Y)|.
\end{aligned}$$

Then, we can further approximate each of the two terms on the upper bound separately. For the first term, using the bound in the work (Wang et al., 2021), we obtain the following upper bound:

$$\begin{aligned}
&|I(T^v; Y) - I(T^{v\prime}; Y)|\\
&\leq B_0^v + B_1^v\sqrt{\mathcal{T}^v}\left(I(X^v; T^v)\right)^{1/2} + B_2^v\left|\mathcal{T}^v\right|^{3/4}\left(I(X^v; T^v)\right)^{1/4}\\
&\quad + B_3^v\sqrt{|\mathcal{T}^v|}\left(I(X^{v\prime}; T^{v\prime})\right)^{1/2} + B_4^v\left|\mathcal{T}^v\right|^{3/4}\left(I(X^{v\prime}; T^{v\prime})\right)^{1/4},
\end{aligned}$$

where $\mathscr{T}^v$ is the finite support of T^v and $T^{v'}$, and B_0^v, B_1^v, B_2^v, B_3^v, and B_4^v are constants that depend on the sequence length K, δ, and $p(x^v)$. Analogously, the second term can be bounded by:

$$
\begin{aligned}
&|I(T^l;Y) - I(T^{l'};Y)| \\
&\leq B_0^l + B_1^l\sqrt{\mathscr{T}^l}\left(I(X^l;T^l)\right)^{1/2} + B_2^l\left|\mathscr{T}^l\right|^{3/4}\left(I(X^l;T^l)\right)^{1/4} \\
&\quad + B_3^l\sqrt{|\mathscr{T}^l|}\left(I(X^{l'};T^{l'})\right)^{1/2} + B_4^l\left|\mathscr{T}^l\right|^{3/4}\left(I(X^{l'};T^{l'})\right)^{1/4},
\end{aligned}
$$

where $\mathscr{T}^l$ is the finite support of T^l and $T^{l'}$, and B_0^l, B_1^l, B_2^l, B_3^l, and B_4^l are constants that depend on the sequence length L, δ, and $p(x^l)$. Combining the above two terms, Eq. (11) is proved. □

References

Agarwal, V., Shetty, R., & Fritz, M. (2020). Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *IEEE conference on computer vision and pattern recognition* (pp. 9690–9698).

Agrawal, A., Kajić, I., Bugliarello, E., Davoodi, E., Gergely, A., Blunsom, P., & Nematzadeh, A. (2022). Rethinking evaluation practices in visual question answering: A case study on out-of-distribution generalization. arXiv preprint arXiv:2205.12191.

Ahuja, K., Caballero, E., Zhang, D., Bengio, Y., Mitliagkas, I., & Rish, I. (2021). Invariance principle meets information bottleneck for out-of-distribution generalization. In *Neural information processing systems* (pp. 3438–3450).

Alayrac, J. B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., & Reynolds, M., et al. (2022). Flamingo: A visual language model for few-shot learning. In *Neural information processing systems* (pp. 23716–23736).

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In *IEEE conference on computer vision and pattern recognition* (pp. 6077–6086).

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015), Vqa: Visual question answering. In *IEEE conference on computer vision and pattern recognition* (pp. 2425–2433)

Ban, Y., & Dong, Y. (2022). Pre-trained adversarial perturbations. In *Neural information processing systems* (pp. 1196–1209).

Bao, F. (2021). Disentangled variational information bottleneck for multiview representation learning. In *International conference on artificial intelligence* (pp. 91–102).

Barber, D., & Agakov, F. (2003). The im algorithm: A variational approach to information maximization. In *Neural information processing systems* (pp. 201–208).

Belghazi, M. I., Baratin, A., Rajeswar, S., Ozair, S., Bengio, Y., Courville, A., & Hjelm, R. D. (2018). Mutual information neural estimation. *International Conference on Machine Learning, 80*, 530–539.

Bennasar, M., Hicks, Y., & Setchi, R. (2015). Feature selection using joint mutual information maximisation. *Expert Systems with Applications, 42*(22), 8520–8532.

Ben-Younes, H., Cadene, R., Thome, N., & Cord, M. (2019). Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection. *Association for the Advancement of Artificial Intelligence, 33*, 8102–8109.

Cadene, R., Dancette, C., Cord, M., Parikh, D., et al. (2019). RUBi: Reducing unimodal biases for visual question answering. In *Neural information processing systems* (pp. 841–852).

Changpinyo, S., Sharma, P., Ding, N., & Soricut, R. (2021). Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *IEEE conference on computer vision and pattern recognition* (pp. 3558–3568).

Chen, X., Fang, H., Lin, T. Y., Vedantam, R., Gupta, S., Dollár, P., & Zitnick, C. L. (2015). Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325.

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

Chen, L., Yan, X., Xiao, J., Zhang, H., Pu, S., & Zhuang, Y. (2020). Counterfactual samples synthesizing for robust visual question answering. In *IEEE conference on computer vision and pattern recognition* (pp. 10800–10809).

Cheng, P., Hao, W., Dai, S., Liu, J., Gan, Z., & Carin, L. (2020). CLUB: A contrastive log-ratio upper bound of mutual information. *International Conference on Machine Learning, 119*, 1779–1788.

Cho, J., Lei, J., Tan, H., & Bansal, M. (2021). Unifying vision-and-language tasks via text generation. *International Conference on Machine Learning, 139*, 1931–1942.

Clark, C., Yatskar, M., & Zettlemoyer, L. (2019). Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Conference on empirical methods in natural language processing* (pp. 4067–4080).

Dancette, C., Cadene, R., Teney, D., & Cord, M. (2021). Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering. In *IEEE international conference on computer vision* (pp. 1574–1583).

Dong, X., Luu, A. T., Lin, M., Yan, S., & Zhang, H. (2021). How should pre-trained language models be fine-tuned towards adversarial robustness? In *Neural information processing systems* (pp. 4356–4369).

Dou, Z. Y., Xu, Y., Gan, Z., Wang, J., Wang, S., Wang, L., Zhu, C., Zhang, P., Yuan, L., & Peng, N., et al. (2022). An empirical study of training end-to-end vision-and-language transformers. In *IEEE conference on computer vision and pattern recognition* (pp. 18166–18176).

Du, Y., Xu, J., Xiong, H., Qiu, Q., Zhen, X., Snoek, C. G., & Shao, L. (2020). Learning to learn with variational information bottleneck for domain generalization. In *European conference on computer vision* (pp. 200–216).

Dubois, Y., Kiela, D., Schwab, D. J., & Vedantam, R. (2020). Learning optimal representations with the decodable information bottleneck. In *Neural information processing systems* (pp. 18674–18690).

Federici, M., Dutta, A., Forré, P., Kushman, N., & Akata, Z. (2020). Learning robust representations via multi-view information bottleneck. In *International conference on learning representations*.

Gan, Z., Chen, Y.C., Li, L., Zhu, C., Cheng, Y., & Liu, J. (2020). Large-scale adversarial training for vision-and-language representation learning. In *Neural information processing systems* (pp. 6616–6628).

Gat, I., Schwartz, I., Schwing, A., & Hazan, T. (2020). Removing bias in multi-modal classifiers: Regularization by maximizing functional entropies. In *Neural information processing systems* (pp. 3197–3208).

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., & Parikh, D. (2017). Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *IEEE conference on computer vision and pattern recognition* (pp. 6904–6913).

Hu, R., Andreas, J., Darrell, T., & Saenko, K. (2018). Explainable neural computation via stack neural module networks. In *European conference on computer vision* (pp. 53–69).

Hu, R., Singh, A., Darrell, T., & Rohrbach, M. (2020). Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In *IEEE conference on computer vision and pattern recognition* (pp. 9992–10002).

Huang, Z., Zeng, Z., Huang, Y., Liu, B., Fu, D., & Fu, J. (2021). Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *IEEE conference on computer vision and pattern recognition* (pp. 12976–12985).

Huang, Z., Zeng, Z., Liu, B., Fu, D., & Fu, J. (2020). Pixel-BERT: Aligning image pixels with text by deep multi-modal transformers. arXiv preprint arXiv:2004.00849.

Hudson, D. A., & Manning, C. D. (2019). Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE conference on computer vision and pattern recognition* (pp. 6700–6709).

Jeon, I., Lee, W., Pyeon, M., & Kim, G. (2021). Ib-gan: Disengangled representation learning with information bottleneck generative adversarial networks. *Association for the Advancement of Artificial Intelligence, 35*, 7926–7934.

Jiang, J., Liu, Z., Liu, Y., Nan, Z., & Zheng, N. (2021). X-ggm: Graph generative modeling for out-of-distribution generalization in visual question answering. In *ACM international conference on multimedia* (pp. 199–208).

Jiang, Y., Natarajan, V., Chen, X., Rohrbach, M., Batra, D., & Parikh, D. (2018). Pythia v0. 1: The winning entry to the vqa challenge 2018. arXiv preprint arXiv:1807.09956.

Kant, Y., Moudgil, A., Batra, D., Parikh, D. & Agrawal, H. (2021). Contrast and classify: Training robust vqa models. In *IEEE international conference on computer vision* (pp. 1604–1613).

Kazemi, V., & Elqursh, A. (2017). Show, ask, attend, and answer: A strong baseline for visual question answering. arXiv preprint arXiv:1704.03162.

Kervadec, C., Antipov, G., Baccouche, M., & Wolf, C. (2021). Roses are red, violets are blue...but should vqa expect them to? In *IEEE conference on computer vision and pattern recognition* (pp. 2776–2785).

Kim, J. H., Jun, J., & Zhang, B. T. (2018). Bilinear attention networks. In *Neural information processing systems* (pp. 1564–1574).

Kim, W., Son, B., & Kim, I. (2021). Vilt: Vision-and-language transformer without convolution or region supervision. *International Conference on Machine Learning, 139*, 5583–5594.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L. J., Shamma, D. A., et al. (2017). Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision, 123*(1), 32–73.

Li, L., Gan, Z., & Liu, J. (2020). A closer look at the robustness of vision-and-language pre-trained models. arXiv preprint arXiv:2012.08673.

Li, L., Lei, J., Gan, Z., & Liu, J. (2021). Adversarial vqa: A new benchmark for evaluating the robustness of vqa models. In *IEEE conference on computer vision and pattern recognition* (pp. 2042–2051).

Li, J., Selvaraju, R. R., Gotmare, A., Joty, S. R., Xiong, C., & Hoi, S. C. (2021). Align before fuse: Vision and language representation learning with momentum distillation. In *Neural information processing systems* (pp. 9694–9705).

Li, C., Xu, H., Tian, J., Wang, W., Yan, M., Bi, B., Ye, J., Chen, H., Xu, G., Cao, Z., et al. (2022). mplug: Effective and efficient vision-language learning by cross-modal skip-connections. In *Conference on empirical methods in natural language processing* (pp. 7241–7259).

Li, C., Yan, M., Xu, H., Luo, F., Wang, W., Bi, B., & Huang, S. (2021). SemVLP: Vision-language pre-training by aligning semantics at multiple levels. arXiv preprint arXiv:2103.07829.

Li, L. H., Yatskar, M., Yin, D., Hsieh, C. J., & Chang, K. W. (2019). Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557.

Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., & Wei, F., et al. (2020). Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European conference on computer vision* (pp. 121–137).

Li, J., Li, D., Xiong, C., & Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *International Conference on Machine Learning, 162*, 12888–12900.

Li, Y., Pan, Y., Yao, T., Chen, J., & Mei, T. (2021). Scheduled sampling in vision-language pretraining with decoupled encoder-decoder network. *Association for the Advancement of Artificial Intelligence, 35*, 8518–8526.

Li, B., Shen, Y., Wang, Y., Zhu, W., Li, D., Keutzer, K., & Zhao, H. (2022). Invariant information bottleneck for domain generalization. *Association for the Advancement of Artificial Intelligence, 36*, 7399–7407.

Liu, X., Li, L., Wang, S., Zha, Z. J., Meng, D., & Huang, Q. (2019). Adaptive reconstruction network for weakly supervised referring expression grounding. In *IEEE international conference on computer vision* (pp. 2611–2620).

Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Neural information processing systems* (pp. 13–23).

Lu, J., Goswami, V., Rohrbach, M., Parikh, D., & Lee, S. (2020). 12-in-1: Multi-task vision and language representation learning. In *IEEE conference on computer vision and pattern recognition* (pp. 10437–10446).

Lu, J., Lin, X., Batra, D., & Parikh, D. (2015). Deeper lstm and normalized cnn visual question answering model. https://github.com/VT-vision-lab/VQA_LSTM_CNN.

Mahabadi, R. K., Belinkov, Y., & Henderson, J. (2021). Variational information bottleneck for effective low-resource fine-tuning. In *International conference on learning representations*.

Nam, J., Cha, H., Ahn, S. S., Lee, J., & Shin, J. (2020). Learning from failure: De-biasing classifier from biased classifier. In *Neural information processing systems* (pp. 20673–20684).

Nguyen, X., Wainwright, M. J., & Jordan, M. I. (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory, 56*(11), 5847–5861.

Oord, Avd, Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748.

Ordonez, V., Kulkarni, G., & Berg, T. (2011). Im2text: Describing images using 1 million captioned photographs. In *Neural information processing systems* (pp. 1143–1151).

Pan, Y., Li, Z., Zhang, L., & Tang, J. (2022). Causal inference with knowledge distilling and curriculum learning for unbiased vqa. *ACM Transactions on Multimedia Computing, Communications, and Applications, 18*(3), 1–23.

Pan, Z., Niu, L., Zhang, J., & Zhang, L. (2021). Disentangled information bottleneck. *Association for the Advancement of Artificial Intelligence, 35*, 9285–9293.

Poole, B., Ozair, S., Van Den Oord, A., Alemi, A., & Tucker, G. (2019). On variational bounds of mutual information. *International Conference on Machine Learning, 97*, 5171–5180.

Shah, M., Chen, X., Rohrbach, M., & Parikh, D. (2019). Cycle-consistency for robust visual question answering. In *IEEE conference on computer vision and pattern recognition* (pp. 6649–6658).

Sharma, P., Ding, N., Goodman, S., & Soricut, R. (2018). Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Annual meeting of the association for computational linguistics* (pp. 2556–2565).

Sheng, S., Singh, A., Goswami, V., Magana, J. A. L., Galuba, W., Parikh, D., & Kiela, D. (2021). Human-adversarial visual question answering. In *Neural information processing systems* (pp. 20346–20359).

Shi, L., Shuang, K., Geng, S., Su, P., Jiang, Z., Gao, P., Fu, Z., de Melo, G., & Su, S. (2020.) Contrastive visual-linguistic pretraining. arXiv preprint arXiv:2007.13135.

Shi, J., Zhang, H., & Li, J. (2019). Explainable and explicit visual reasoning over scene graphs. In *IEEE conference on computer vision and pattern recognition* (pp. 8376–8384).

Shrestha, R., Kafle, K., & Kanan, C. (2020). A negative case analysis of visual grounding methods for vqa. In *Annual meeting of the association for computational linguistics* (pp. 8172–8181).

Shwartz-Ziv, R., & Tishby, N. (2017). Opening the black box of deep neural networks via information. arXiv preprint arXiv:1703.00810.

Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei. F., & Dai, J. (2020). VL-BERT: Pre-training of generic visual-linguistic representations. In *International conference on learning representations*.

Sun, S., Chen, Y. C., Li, L., Wang, S., Fang, Y., & Liu, J. (2021). Lightningdot: Pre-training visual-semantic embeddings for real-time image-text retrieval. In *Annual meeting of the association for computational linguistics* (pp. 982–997).

Tan, H., & Bansal, M. (2019). LXMERT: Learning cross-modality encoder representations from transformers. In *Conference on empirical methods in natural language processing* (pp. 5099–5110).

Teney, D., Kafle, K., Shrestha, R., Abbasnejad, E., Kanan, C., & Hengel, A. V. D. (2020). On the value of out-of-distribution testing: An example of goodhart's law. In *Neural information processing systems* (pp. 407–417).

Tishby, N., & Zaslavsky, N. (2015). Deep learning and the information bottleneck principle. In *IEEE information theory workshop* (pp. 1–5).

Tishby, N., Pereira, F. C., & Bialek, W. (2000). The information bottleneck method. arXiv preprint physics/0004057.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Neural information processing systems* (pp. 5998–6008).

Wang, W., Bao, H., Dong, L., & Wei, F. (2021). Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. arXiv preprint arXiv:2111.02358

Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O. K., Singhal, S., Som, S. et al. (2023). Image as a foreign language: Beit pretraining for all vision and vision-language tasks. In *IEEE conference on computer vision and pattern recognition* (pp. 19175–19186)

Wang, H., Guo, X., Deng, ZH., & Lu, Y. (2022). Rethinking minimal sufficient representation in contrastive learning. In *IEEE conference on computer vision and pattern recognition* (pp. 16041–16050).

Wang, B., Wang, S., Cheng, Y., Gan, Z., Jia, R., Li, B., & Liu, J. (2021). InfoBERT: Improving robustness of language models from an information theoretic perspective. In *International conference on learning representations*.

Wang, Z., Yu, J., Yu, A. W., Dai, Z., Tsvetkov, Y., & Cao, Y. (2022). Simvlm: Simple visual language model pretraining with weak supervision. In *International conference on learning representations*

Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., & Yang, H. (2022). OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *International Conference on Machine Learning, 162*, 23318–23340.

Whitehead, S., Wu, H., Fung, Y.R., Ji, H., Feris, R., & Saenko, K. (2020). Learning from lexical perturbations for consistent visual question answering. arXiv preprint arXiv:2011.13406.

Xu, H., Ye, Q., Yan, M., Shi, Y., Ye, J., Xu, Y., Li, C., Bi, B., Qian, Q., & Wang, W., et al. (2023). mplug-2: A modularized multi-modal foundation model across text, image and video. arXiv preprint arXiv:2302.00402.

Yang, Z., He, X., Gao, J., Deng, L., & Smola, A. (2016). Stacked attention networks for image question answering. In *IEEE conference on computer vision and pattern recognition* (pp. 21–29).

Yu, L., Poirson, P., Yang, S., Berg, A. C., & Berg, T. L. (2016). Modeling context in referring expressions. In *European conference on computer vision* (pp. 69–85).

Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., & Wu, Y. (2022). Coca: Contrastive captioners are image-text foundation models. arXiv preprint arXiv:2205.01917

Yuan, L., Chen, D., Chen, Y. L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., et al. (2021). Florence: A new foundation model for computer vision. arXiv preprint arXiv:2111.11432

Yu, F., Tang, J., Yin, W., Sun, Y., Tian, H., Wu, H., & Wang, H. (2021). Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. *Association for the Advancement of Artificial Intelligence, 35*, 3208–3216.

Zeng, Y., Zhang, X., Li, H., Wang, J., Zhang, J., & Zhou, W. (2022). X^2-VLM: All-in-one pre-trained model for vision-language tasks. arXiv preprint arXiv:2211.12402

Zeng, Y., Zhang, X., & Li, H. (2022). Multi-grained vision language pre-training: Aligning texts with visual concepts. *International Conference on Machine Learning, 162*, 25994–26009.

Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., & Gao, J. (2021). Vinvl: Revisiting visual representations in vision-language models. In *IEEE conference on computer vision and pattern recognition* (pp. 5579–5588).

Zhang, Z., Zhao, Z., Lin, Z., He, X., et al. (2020). Counterfactual contrastive learning for weakly-supervised vision-language grounding. In *Neural information processing systems* (pp. 18123–18134).

Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L. H., Zhou, L., Dai, X., Yuan, L., & Li, Y., et al. (2022). Regionclip: Region-based language-image pretraining. In *IEEE conference on computer vision and pattern recognition* (pp. 16793–16803).

Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J., & Gao, J. (2020). Unified vision-language pre-training for image captioning and vqa. *Association for the Advancement of Artificial Intelligence, 34*, 13041–13049.

Zhou, D., Yu, Z., Xie, E., Xiao, C., Anandkumar, A., Feng, J., & Alvarez, J. M. (2022). Understanding the robustness in vision transformers. *International Conference on Machine Learning, 162*, 27378–27394.

Zhu, Y., Groth, O., Bernstein, M., & Fei-Fei, L. (2016). Visual7w: Grounded question answering in images. In *IEEE conference on computer vision and pattern recognition* (pp. 4995–5004).