

A. Supplementary Material

In this supplementary document, we present additional material about our CIIC model from the following aspects:

A.1. Formula Derivations. **A.2.** Additional Ablation Experiments. **A.3.** Additional quantitative analysis on deconfounding. **A.4.** Qualitative Results and Visualization. **A.5.** Limitations.

A.1. Formula Derivations

Derivation of Causal Intervention $P(Y|do(X))$. In our main paper, we respectively apply the Normalized Weighted Geometric Mean (NWGM) approximation [41] to compute the Eq. (2) in Section 3.1 and Eq. (8) in Section 3.3. In this section, we give the detailed derivation of Eq. (3) and Eq. (9). Before showing how to use NWGM to moving the outer expectation into the Softmax in Eq. (3) and Eq. (9), we first introduce the definition of Weighted Geometric Mean (WGM) of a function $y(x)$ as follows [41]:

$$WGM(y(x)) = \prod_x y(x)^{P(x)}, \quad (12)$$

where the exponential term $P(x)$ represents the probability distribution of x . If $y(x)$ is an exponential function, *i.e.*, $y(x) = \exp[f(x)]$, Eq. (12) can be rewritten as [41]:

$$\begin{aligned} WGM(y(x)) &= \prod_x y(x)^{P(x)} \\ &= \prod_x \exp[f(x)]^{P(x)} \\ &= \prod_x \exp[f(x)P(x)] \\ &= \exp\left[\sum_x f(x)P(x)\right] \\ &= \exp\{\mathbb{E}_x[f(x)]\}, \end{aligned} \quad (13)$$

where the expectation \mathbb{E}_x is absorbed into the exponential term. Consequently, the expectation of $y(x)$ can be approximated as follows:

$$\begin{aligned} \mathbb{E}_x[y(x)] &= \sum_x y(x)P(x) \\ &\approx WGM(y(x)) = \exp\{\mathbb{E}_x[f(x)]\}, \end{aligned} \quad (14)$$

where $y(x) = \exp[f(x)]$. Thus, the Normalized Weighted Geometric Mean (NWGM) approximation is defined as [41]:

$$\begin{aligned} NWGM(y(x)) &= \frac{\prod_x \exp(f(x))^{P(x)}}{\sum_j \prod_x \exp(f(x))^{P(x)}} \\ &= \frac{\exp(\mathbb{E}_x[f(x)])}{\sum_j \exp(\mathbb{E}_x[f(x)])} \\ &= \text{Softmax}(\mathbb{E}_x[f(x)]). \end{aligned} \quad (15)$$

In our proposed IOD, since $P(Y|do(X))$ (Eq. (2) of the submitted manuscript) is used as the predictive function of the class label, it is natural to parameterize it as a network with a Softmax layer as the last layer. We have

$$P(Y|X, Z) = \text{Softmax}[g(X, Z)] \propto \exp[g(X, Z)]. \quad (16)$$

According to Eq. (2) of the manuscript and Eq. (16), we have

$$\begin{aligned} P(Y|do(X = x)) &= \sum_Z P(Y|X = x, Z = z)P(Z = z) \\ &= \mathbb{E}_{[Z]}[P(Y|Z = z, X = x)] \\ &\approx WGM(P(Y|Z = z, X = x)) \\ &\approx \exp\{[g(\mathbb{E}_Z[Z], x)]\}. \end{aligned} \quad (17)$$

To guarantee the sum of $p(Y|do(x))$ to be 1, Eq. (17) can be further normalized according to Eq. (15) as:

$$P(Y|do(X = x)) \approx \text{softmax}[g(\mathbb{E}_Z[Z], x)]. \quad (18)$$

Given X 's RoI feature \mathbf{x} whose class label is y^c , we parameterize $P(y^c|do(X = \mathbf{x}))$ as a network to introduce causal intervention into the classifier. The last layer of this network for class prediction is the Softmax layer that implements $P(y^c|do(X = \mathbf{x}))$ as:

$$\begin{aligned} P(y^c|do(X = \mathbf{x})) \\ \approx \text{Softmax}(\mathbf{W}_1 \mathbf{x} + \mathbf{W}_2 \cdot \mathbb{E}_z[g_x(\mathbf{z})]), \end{aligned} \quad (19)$$

where $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{n \times d}$ denote the learnable weight matrices. Note that we set \mathbf{z} to be conditioned on \mathbf{x} since, if not, the expectation of \mathbf{z} will degrade to a fixed vector. As validated in [21, 22], this trick can effectively increase the representation power of the whole model. Assume a fixed confounder dictionary $Z = [\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_n]$, where n is the class size in dataset and $\mathbf{z}_i \in \mathbb{R}^d$ is the average RoI feature $\bar{\mathbf{x}}_i$ of objects in i -th class, and $P(\mathbf{z}_i) = 1/n$, *e.g.*, a uniform prior of each object class, we have:

$$\mathbb{E}_z[g_x(\mathbf{z})] = \frac{1}{n} \sum_{i=1}^n P(y_i^c|\mathbf{x}) \mathbf{z}_i, \quad (20)$$

where y_i^c is the i -th class label and $P(y_i^c|\mathbf{x})$ is the pre-trained classifier's probability output that \mathbf{x} belongs to class y_i^c . Following Eq. (18), Eq. (19) and Eq. (20), we have:

$$\begin{aligned} P(Y|do(X = \mathbf{x})) \\ \approx P\left(Y|\text{concat}\left(\mathbf{x}, \frac{1}{n} \sum_{i=1}^n P(y_i^c|\mathbf{x}) \mathbf{z}_i\right)\right), \end{aligned} \quad (21)$$

which is the training objective of the proposed IOD and given in Eq. (3) of our main paper.

Derivation of Causal Intervention $P(W|do(V), do(\mathbf{h}_1))$. Given one attended visual feature \mathbf{v} and its corresponding word w , $P(w|\mathbf{v}, \mathbf{h}_1, D_1, D_2)$ can be first parameterized as a network to incorporate causal intervention into the Transformer decoder. If the last layer of this network for word prediction is the Softmax layer, similar to Eq. (18), $P(W|do(V), do(\mathbf{h}_1))$ can be formulated as follows:

$$\begin{aligned} P(W = w|do(V = \mathbf{v}), do(\mathbf{h}_1)) \\ \approx \text{Softmax}(\mathbb{E}_{D_1} \mathbb{E}_{D_2}[g(\mathbf{v}, \mathbf{h}_1, D_1, D_2)]), \end{aligned} \quad (22)$$

where $g(\cdot)$ denotes the linear embedding layer before the Softmax layer.

In particular, if the linear model $g(\mathbf{v}, \mathbf{h}_1, D_1, D_2) = \mathbf{Q}_1 \cdot$

$\mathbf{h}_2 + \mathbf{Q}_2 \cdot D_1 + \mathbf{Q}_3 \cdot D_2$, where $\mathbf{Q}_1, \mathbf{Q}_2, \mathbf{Q}_3 \in \mathbb{R}^{n \times d}$ denote the matrices of learnable weights and $\mathbf{h}_2 = \mathbf{v} + \mathbf{h}_1$, we have:

$$P(W = w | do(V = \mathbf{v}), do(\mathbf{h}_1)) \approx \text{Softmax}\{g(\mathbf{h}_2, \mathbb{E}_{D_1}[D_1], \mathbb{E}_{D_2}[D_2])\}, \quad (23)$$

which is given in Eq. (9) of our main paper.

Likewise, we set D_1 and D_2 to be conditioned on the fused feature \mathbf{h}_2 to increase the representation power of the whole model as follows:

$$P(W = w | do(V = \mathbf{v}), do(\mathbf{h}_1)) \approx \text{Softmax}\{g(\mathbf{h}_2, \mathbb{E}_{[D_1|\mathbf{h}_2]}[D_1], \mathbb{E}_{[D_2|\mathbf{h}_2]}[D_2])\}. \quad (24)$$

In our main paper, we build the approximate visual confounder dictionary \mathbb{D}_1 and linguistic confounder dictionary \mathbb{D}_2 . Thus, $\mathbb{E}_{[D_1|\mathbf{h}_2]}[D_1]$ and $\mathbb{E}_{[D_2|\mathbf{h}_2]}[D_2]$ can be computed as follows:

$$\mathbb{E}_{[D_1|\mathbf{h}_2]}[D_1] = \text{Softmax}(\mathbb{D}_1 \mathbf{h}_2) \mathbb{D}_1. \quad (25)$$

$$\mathbb{E}_{[D_2|\mathbf{h}_2]}[D_2] = \text{Softmax}(\mathbb{D}_2 \mathbf{h}_2) \mathbb{D}_2. \quad (26)$$

A.2. Additional Ablation Experiments

Encoder and Decoder Layers. To investigate the impact of the number of encoder and decoder layers, we perform Base+Glove+CI with different numbers of the stacked blocks $L \in \{1, 2, 4, 6\}$. Table 5 reports the performance of Base+GloVe+CI using different L . We can see that as L increases, its performance gradually improves and reaches the optimal value when $L = 6$. This is due to the fact that deeper layers encourage the encoder of captioner to represent more complicated relationships between objects, and the decoder to provide more discriminative latent vectors for the prediction of words.

Table 5. The performance of Base+GloVe+CI with different numbers of attention blocks $L \in \{1, 2, 4, 6\}$. The results are reported after the XE training stage.

L	Cross-Entropy Loss							
	B@1	B@2	B@3	B@4	M	R	C	S
1	75.9	59.9	46.7	36.0	28.0	56.5	114.1	21.0
2	76.2	60.3	46.9	36.3	28.1	56.9	116.5	21.2
4	76.3	60.6	47.1	36.3	28.3	57.0	117.0	21.2
6	76.5	60.8	47.1	36.5	28.4	57.0	117.1	21.3

Effect of the IOD Features. To evaluate the effectiveness of our IOD features, we perform three popular used models, *i.e.*, Up-Down [2], AoANet [16] and Transformer, with ablative features in our experiment. In order to apply both the IOD and bottom-up features to each model, we align them by extracting the IOD features with the same bounding box coordinates of Up-Down [2]. Table 6 shows the performance of three representative models with different ablative features. Specifically, for each model, we use the following four ablative feature settings: 1) **Bottom-up**: the widely-used bottom-up features from Up-Down [2];

Table 6. The performance of three representative image captioning models with ablative features on Karpathy split. All results are reported after the SCST optimization stage.

Model	Feature	B@4	M	R	C
Up-Down [2]	Bottom-up	36.3	27.7	56.9	120.1
	Only IOD	34.4	27.2	56.6	116.3
	+Ent	37.5	28.0	58.3	125.9
	+IOD	39.0	28.8	58.8	129.5
AoANet [16]	Bottom-up	38.9	29.2	58.8	129.8
	Only IOD	35.5	27.5	56.8	119.1
	+Ent	39.0	28.9	58.7	130.6
	+IOD	39.3	29.2	58.8	130.8
Transformer	Bottom-up	38.4	28.6	58.4	128.6
	Only IOD	35.5	27.3	56.8	120.7
	+Ent	38.8	28.9	58.7	130.3
	+IOD	39.1	29.2	59.1	131.0

2) **Only IOD**: pure IOD features; 3) **+Ent**: the entangled features from training the IOD without causal intervention. “+” denotes the extracted features are concatenated with the bottom-up features; 4) **+IOD**: the disentangled IOD features with causal intervention, concatenated to the bottom-up features. From Table 6, we can see that by means of our +IOD trained on MS-COCO, each model can achieve absolute gains over most of the metrics. In particular, the Up-Down model with our +IOD achieves a huge performance improvement in comparison with +Bottom-up (from 120.1 CIDEr score to 129.5 CIDEr score). When comparing +IOD with +Ent without intervention, each model achieves superior performances over all metrics, which validates the effectiveness of our IOD features based on causal intervention. We can also observe that only exploiting the pure IOD features (*i.e.*, Only IOD) would hurt the model performance. The reason is that other than the IOD features, the bottom-up features contain the additional attribute information, which contributes to the generation of detailed captions.

Table 7. Training burdens and the bias degree of different models on Karpathy split. A@Gen/A@Act represents the average accuracy of gender/action words to evaluate the gender/action bias.

Model	GPU Hours	Time(sec.)/Batch	A@Gen	A@Act
Transformer	2.7	0.085	0.620	0.690
Transformer+ITD	3.4	0.105	0.635	0.718
Transformer+IOD	3.7	0.116	0.662	0.742
CIIC _G	4.2	0.135	0.684	0.753

Training burdens. Table 7 shows the training burdens (GPU hours per 10 epochs and training time per batch with a batch size of 10) of CIIC, CI and non-CI baselines on one 3080 GPU. CIIC need more computational costs (4.2/2.7=1.5) since causal intervention is introduced into both the encoder and decoder.

A.3. Additional quantitative analysis on deconfounding

Table 7 also reports experimental results of different models on the gender/action bias after the XE training stage. Specifically, we calculate whether the gender words (e.g. “man”, “woman”, “girl” and “boy”) or action words (e.g. “eat”, “ride” and “hold”) are consistent between the generated caption and the ground truth for 80 visual objects. From Table 7, we can see that the accuracies of gender and action are respectively improved by 10.3% and 9.1% when IOD and ITD are used in Transformer. In addition, we train CIIC on the MSCOCO-Bias dataset [14] using the XE loss. CIIC still outperforms the transformer baseline significantly (error rate: 0.085 vs. 0.125, gender ratio: 0.426 vs. 0.280), which further confirms the effectiveness of CIIC on the biased dataset.

A.4. Qualitative Results and Visualization

To further validate the effectiveness of our CIIC model, we complement the additional qualitative analysis and visualization experiments. Figure 8 illustrates the additional captions generated by CIIC and the original Transformer. From Figure 8, it can be seen that the original transformer model generates logically right captions, but these captions might not be consistent with the image contents. In contrast, our proposed CIIC is able to generate more grounded and reasonable descriptions. Specifically, our CIIC shows more superior performance over the transformer baseline. Firstly, CIIC is able to specify the number of objects of the same kind more precisely. For example, there are many signs in the image of the third row (left). But, the Transformer baseline only finds one sign while CIIC is able to count correctly. Secondly, CIIC describes the interactions of objects in an image more accurately. In the image of the third row (right), CIIC can identify that the motorcycles are parking on the side of a street but not a building. In addition, in the image of the last row (right), a man is milking a cow not just standing. These advantages of CIIC mainly result from the implementation of causal intervention. In the encoder, by virtue of self-attention and the object features disentangled by the proposed IOD, our CIIC is able to more accurately represent the relationships among the objects of an image. In the decoder, CIIC utilizes causal intervention to effectively avoid the spurious correlations between irrelevant objects.

To qualitatively analyze the effect of causal intervention on the generated captions, we visualize the attended image regions during the caption generation in Figure 9. Observing the attended image regions in Figure 9, we find that our CIIC model is able to correctly ground image regions to the words, while the Transformer baseline attends to unreasonable regions and then generates incorrect captions, e.g., the Transformer model attends to the cake region and pre-

dict the word “fork” caused by the dataset bias. In contrast, our CIIC can effectively suppress the dataset bias and accurately attends to the spoon region so as to generate the word “spoon”.

A.5. Limitations

Although we leverage causal intervention to disentangle the region-based features and deconfound the image captioning to generate more grounded captions, the visual and linguistic confounders still can not be completely deconfounded in the experiments. Four examples of generating inconsistent captions are shown in Figure 10. We can see that though our CIIC model can effectively alleviate the spurious correlations in the case of dataset bias, it still generates the biased sentences in practice. The limitations of our method are as follows: First, the confounder dictionaries in our CIIC model are approximately built, which makes it difficult to utilize the exact confounder to fully eliminate the spurious correlations. Second, as illustrated in Figure 9, our model does not distinguish visual words and non-visual words at each time step, which is not good for a more fine-grained captioning generation. Third, the SCST optimization of CIIC may cause biases, which is in contradiction to causal intervention. To overcome these limitations, we will continue our future works in three directions. First, we will explore how to further enhance the effect of causal intervention, for example adaptively learning the confounder dictionaries. Second, we will investigate certain adaptive attention mechanism to dynamically measure the contributions of visual and language cues at each decoding time step. Last but not the least, we will combine our proposed CIIC with the image-text matching or visual grounding models to improve the grounding performance of our method further.









	<p>GT: The street signs for Gladys and Detroit streets are attached to a wooden pole.</p> <p>Transformer: Two street signs on top of a pole.</p> <p>CIIC: Two street signs on the side of a pole.</p>		<p>GT: A woman sits on a luggage case on a sidewalk.</p> <p>Transformer: A woman standing next to two suitcases.</p> <p>CIIC: A woman sitting on top of a suitcase.</p>
	<p>GT: A piece of cake on a plate with some juice by it.</p> <p>Transformer: A piece of cake on a plate on a table.</p> <p>CIIC: A piece of cake on a plate with a glass of orange juice.</p>		<p>GT: Bathroom with a shower, sink, and toilet in it.</p> <p>Transformer: A bathroom with a toilet and a shower.</p> <p>CIIC: A bathroom with a sink and a toilet and a shower .</p>
	<p>GT: Many different signs cover a post in front of a building.</p> <p>Transformer: A street sign in front of a building.</p> <p>CIIC: A group of street signs in front of a building.</p>		<p>GT: A bunch of motorcycles parked along the side of the street.</p> <p>Transformer: A group of motorcycles parked on the side of a building.</p> <p>CIIC: A row of motorcycles parked on the side of a street.</p>
	<p>GT: A cat that is sitting on a bed next to a book.</p> <p>Transformer: An orange cat laying on top of a book.</p> <p>CIIC: An orange cat sitting on a bed next to a book.</p>		<p>GT: A man milking a brown and white cow in barn.</p> <p>Transformer: A man standing next to a cow.</p> <p>CIIC: A man milking a cow in a barn.</p>

Figure 8. Additional examples with images and captions generated by CIIC and Transformer, as well as the ground truth captions.



(a) Transformer: A piece of cake on a plate with a fork.



(b) CIIC: A piece of cake on a plate with a spoon.

Figure 9. Visualization of attention regions for sample captions generated by the proposed CIIC and the Transformer baseline, where we outline the region with the highest attention weight in top-down attention in red. The original Transformer model easily attends to unsuitable regions due to dataset bias while our CIIC is less likely so.




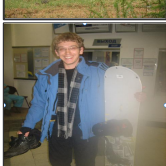
	<p>GT: A woman with lots of tattoos sits on a suitcase in a forest.</p> <p>Transformer: A woman sitting in the grass with a cell phone.</p> <p>CIIC: A woman sitting on a suitcase talking on a cell phone.</p>		<p>GT: Three Zebra's eating grass as they walk.</p> <p>Transformer: Three zebras and other animals grazing in a field.</p> <p>CIIC: Three zebras grazing in the grass in a field.</p>
	<p>GT: Two cows outside one laying down and the other standing near a building.</p> <p>Transformer: Two cows are laying in a field of grass.</p> <p>CIIC: A cow standing in a field with a cows laying down.</p>		<p>GT: A young man holding a snowboard and a pair of shoes.</p> <p>Transformer: A man standing in front of a white refrigerator.</p> <p>CIIC: A man standing in front of a snowboard.</p>

Figure 10. Some failure cases of CIIC. For comparisons, we also show the captions generated by the Transformer baseline.