# Supplementary Material for Probabilistic Debiasing of Scene Graphs

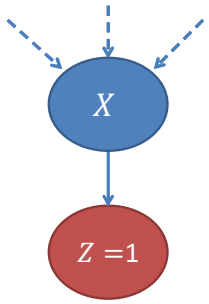
Bashirul Azam Biswas and Qiang Ji
Rensselaer Polytechnic Institute, Troy, NY-12180
{biswab, jiq}@rpi.edu

## A. Virtual evidence

Posterior inference in a Bayesian network entails inferring the values of a set of unknown variables given a set of known variables as evidence. This inference with certain evidence is straightforward (see Sect 2.1 of [4]) whereas inference with uncertain evidence requires a cascaded inference strategy (see Sect 2.2 of [4]). If a random categorical variable $X$ consists of $n$ categories $\{x_1, x_2, ...x_n\}$, then the uncertain evidence of that node can be specified as likelihood ratios $L_{x_1} : L_{x_2} : ... : L_{x_n}$. In order to perform inference with this type of uncertainty, we can insert a virtual evidence node $Z$ as a child of the evidence node $X$ and instantiate that node as *true*. Hence, we achieve the following

$$P(Z = 1|x_1) : P(Z = 1|x_2)... : P(Z = 1|x_n) = L_{x_1} : L_{x_2} : ... : L_{x_n} \tag{13}$$

An illustration of virtual evidence is shown in Figure 7.




$$P(Z = 1|x_1): P(Z = 1|x_2): ...: P(Z = 1|x_n) = L_{x_1}: L_{x_2}: ...: L_{x_n}$$

Figure 7. Virtual evidence node $Z$ as a child of the evidence node $X$.

Now, the observation probability of evidence node $X$ can be achieved from an independent measurement device and we denote this as $P_{obs}(X)$. According to Theorem 5 in [1], we can convert this observation probability to likelihood ratios of the virtual node as follows

$$P(Z = 1|x_1) : P(Z = 1|x_2)... : P(Z = 1|x_n) = \frac{P_{obs}(x_1)}{P(x_1)} : \frac{P_{obs}(x_2)}{P(x_2)} : ... : \frac{P_{obs}(x_n)}{P(x_n)} \tag{14}$$

where $P(X)$ is the marginal probability of node $X$. In our case, we model the triplets with three random categorical variables $S, R$, and $O$, and obtain the observation probabilities of these variables from a trained deep learning-based measurement model $\theta_m$. Hence, we can write

$$
\begin{aligned}
P(Z_s = 1|s_1) : .. : P(Z_s = 1|s_n) &= \frac{P_{I,\theta_m}(s_1)}{P(s_1)} : .. : \frac{P_{I,\theta_m}(s_n)}{P(s_n)} \\
P(Z_o = 1|o_1) : .. : P(Z_o = 1|o_n) &= \frac{P_{I,\theta_m}(o_1)}{P(o_1)} : .. : \frac{P_{I,\theta_m}(o_n)}{P(o_n)} \\
P(Z_r = 1|r_1) : .. : P(Z_r = 1|r_n) &= \frac{P_{I,\theta_m}(r_1)}{P(r_1)} : .. : \frac{P_{I,\theta_m}(r_n)}{P(r_n)}
\end{aligned} \tag{15}
$$

## B. Special cases of MAP inference

In this section, we describe some special cases of MAP inference. For increased readability, we rewrite the MAP inference equation here

$$
\begin{aligned}
S^*, R^*, O^* &= \arg\max_{S,R,O} P(S, R, O|Z_s = 1, Z_o = 1, Z_r = 1) \\
&= \arg\max_{S,R,O} P_{I,\theta_m}(S) P_{I,\theta_m}(O) \frac{P_{I,\theta_m}(R)}{P(R)} P(R|S, O)
\end{aligned} \tag{16}
$$

1. **FREQ baseline [8], subject and object is known**: In this scenario, we predict the relationship label of a triplet only from its known subject ($S = S_g$) and object label ($O = O_g$). Hence, we can write

$$
\begin{aligned}
P_{I,\theta_m}(S) &= \mathbb{1}_{\{S=S_g\}} \\
P_{I,\theta_m}(O) &= \mathbb{1}_{\{O=O_g\}}
\end{aligned} \tag{17}
$$

   Since there is no image measurement for the relationship label, the MAP Eqn. (16) becomes

$$
\begin{aligned}
R^* &= \arg\max_R P(R|S = S_g, O = O_g) \\
&= \arg\max_R P(R|S_g, O_g)
\end{aligned} \tag{18}
$$

   This is essentially the FREQ baseline proposed by [8].

2. **Synthetic relationship, subject, and object is known:** The within-triplet Bayesian network is a generative model for subject, object, and relationship. Hence, we can generate synthetic samples of relationship $R_{syn}$ given subject and object as following

$$
R_{syn} \sim P(R|S_g, O_g) \tag{19}
$$

3. **Uncertain evidence for relationship, subject, and object is known:** In presence of the uncertainty associated only with the relationship label from measurement model $\theta_m$, the MAP decision becomes

$$
\begin{aligned}
R^* &= \arg\max_R P(R|Z_r = 1, S = S_g, O = O_g) \\
&= \arg\max_R \frac{P_{I,\theta_m}(R)}{P(R)} P(R|S_g, O_g)
\end{aligned} \tag{20}
$$

   This is essentially the PredCls setting [7], where we infer the relationship given the categories and bounding boxes of the subject and object of a triplet. Visualization is provided in Figure 8.

4. **No Bayesian network, only evidence uncertainties**: In this case, we infer each component of a triplet independently from its measurement uncertainty. Hence, the MAP decision becomes as follows

$$
\begin{aligned}
S^* &= \arg\max_S P_{I,\theta_m}(S) \\
O^* &= \arg\max_O P_{I,\theta_m}(O) \\
R^* &= \arg\max_R P_{I,\theta_m}(R)
\end{aligned} \tag{21}
$$

   This is essentially predicting the labels of a triplet from the probabilities that the classification head of the SOTA deep learning-based methods produce.

From the discussion above, we can conclude that the MAP inference in Eqn. (16) is the most general framework and all the other cases can be considered as special cases.

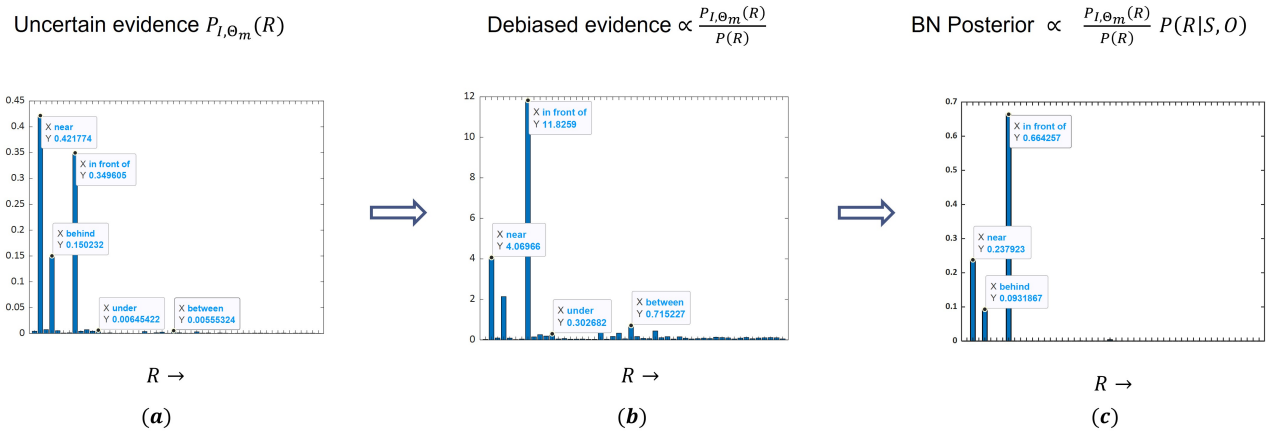



Figure 8. (a) Uncertain evidence of a triplet from VCTree [6] in PredCls setting has the highest support for a majority class 'near'. We also see competing minority counterparts such as 'in front of' and 'behind'; (b) debiasing is increasing the support for minorities. It also amplifies noise by increasing the support for all classes; (c) restoring within-triplet prior by $P(R|S,O)$. The noisy responses in all irrelevant classes are completely removed by the prior. As we observe from (a) to (c), the final MAP inference changes the measured relationship label from a majority class 'near' to a minority class 'in front of'.

## C. Constrained optimization for conflict resolution

In SGCls and SGDet settings, we are required to infer object labels along with relationship labels. Any object entity $E_i$ can reside as subjects or objects in multiple triplets and after MAP inference with Eqn. (16), the inferred values of objects in different triplets may contain different values. We denote $T_i^S$ as the set of triplets where $E_i$ acts as the subject and $T_i^O$ as the set of triplets where $E_i$ acts as the object -

$$\begin{aligned} T_i^S &= \{t_p : t_p(S) = E_i\} \\ T_i^O &= \{t_q : t_q(O) = E_i\} \end{aligned} \tag{22}$$

Now, we define an optimization function by summing up all the posterior probabilities of within-triplet Bayesian network from all triplets in $T_i^S$ and in $T_i^O$ as following

$$f = \sum_{t_p \in T_i^S} P(S_{t_p}, R_{t_p}, O_{t_p} | Z_{S_{t_p}} = 1, Z_{R_{t_p}} = 1, Z_{O_{t_p}} = 1) + \sum_{t_q \in T_i^O} P(S_{t_q}, R_{t_q}, O_{t_q} | Z_{S_{t_q}} = 1, Z_{R_{t_q}} = 1, Z_{O_{t_q}} = 1) \tag{23}$$

Evaluating this optimization function for every possible combination of subjects, relationships, and objects in multiple connected triplets is challenging and hence we propose a two-step alternating approach. In the first step, for each entity, we modify the measurement probabilities of all nodes connected to object entity $E_i$ according to the inferred values. In this way, we essentially reduce all the evidence uncertainty and only keep the uncertainty to object entity $E_i$. The modification can be written as follows

$$\begin{aligned} P_{I,\theta_m}(S_{t_p}) &= P_{I,\theta_m}(O_{t_q}) = P_{I,\theta_m}(E_i) \\ P_{I,\theta_m}(R_{t_p}) &= \mathbb{1}_{\{R_{t_p} = r^I_{t_p}\}}, \quad \forall t_p \in T_i^S \\ P_{I,\theta_m}(O_{t_p}) &= \mathbb{1}_{\{O_{t_p} = o^I_{t_p}\}}, \quad \forall t_p \in T_i^S \\ P_{I,\theta_m}(S_{t_q}) &= \mathbb{1}_{\{S_{t_q} = s^I_{t_q}\}}, \quad \forall t_q \in T_i^O \\ P_{I,\theta_m}(R_{t_q}) &= \mathbb{1}_{\{R_{t_q} = r^I_{t_q}\}}, \quad \forall t_q \in T_i^O \end{aligned} \tag{24}$$

We can simplify the objective function using Eqn. (23) and Eqn. (16) by

$$
\begin{aligned}
f &= \sum_{t_p \in T_i^S} P_{I,\theta_m}(S_{t_p}) P(R = r^I_{t_p} | S_{t_p}, O = o^I_{t_p}) + \sum_{t_q \in T_i^O} P_{I,\theta_m}(O_{t_q}) P(R = r^I_{t_q} | S = s^I_{t_q}, O_{t_q}) \\
&= \sum_{t_p \in T_i^S} P_{I,\theta_m}(E_i) P(R = r^I_{t_p} | E_i, O = o^I_{t_p}) + \sum_{t_q \in T_i^O} P_{I,\theta_m}(E_i) P(R = r^I_{t_q} | S = s^I_{t_q}, E_i)
\end{aligned} \tag{25}
$$

Finally, we can simplify $f(.)$ and infer $E_i$ as

$$
\begin{aligned}
f(E_i) = P_{I,\theta_m}(E_i) \Bigg( & \sum_{t_p \in T_i^S} P(R = r^I_{t_p} | E_i, O = o^I_{t_p}) \\
& + \sum_{t_q \in T_i^O} P(R = r^I_{t_q} | S = s^I_{t_q}, E_i) \Bigg) \\
E_i^* &= \arg\max f(E_i)
\end{aligned} \tag{26}
$$

In the second step, we update the relationship label of each triplet based on the updated subject and object values

$$
\begin{aligned}
R_j^* &= \arg\max_R P(R_j | Z_r = 1, S = s_j, O = O_j) \\
&= \arg\max_R \frac{P_{I,\theta_m}(R_j)}{P(R_j)} P(R_j | S = s_j, O = o_j)
\end{aligned} \tag{27}
$$

We include a pseudo-code for the constrained optimization in Algorithm 1.

**Algorithm 1** Constrained Optimization for conflict resolution

**Require:** Inferred scene graph $\mathcal{G}_I^{Inf}$ from within triplet inference with potential conflicts in objects, maximum iteration $N$
  $\mathcal{G}_I^0 \leftarrow \mathcal{G}_I^{Inf}$
  $n \leftarrow 1$
  **while** $n < N$ **do**
    **for** each $E_i \in \mathcal{G}_I^{n-1}$ **do** ▷ update each object value with constraints
      find triplets where $E_i$ acts as subject and object by Eqn. (22)
      modify uncertain evidence of connected triplets except $E_i$ by Eqn. (24)
      infer $E_i^*$ by Eqn. (26)
    **end for**
    **for** each $R_j \in \mathcal{G}_I^{n-1}$ **do** ▷ update each relationship with updated object values
      infer $R_j^*$ by Eqn. (27)
    **end for**
    Construct new scene graph $\mathcal{G}_I^n = \{\mathcal{E}^*, \mathcal{R}^*\}$
    **if** $\mathcal{G}_I^n == \mathcal{G}_I^{n-1}$ **then**
      return $\mathcal{G}_I^n$
    **else**
      $\mathcal{G}_I^{n-1} = \mathcal{G}_I^n$
      $n = n + 1$
    **end if**
  **end while**
**Ensure:** Inferred scene graph $\mathcal{G}_I^n$ with no conflict in objects

## D. Pseudo-code for sample augmentation

We include a pseudo-code for sample augmentation in Algorithm 2. The concatenation of triplet entities is done by simply adding the class labels by a 'space' to make it a valid sentence that can be fed into a sentence-embedding model to generate embeddings.

**Algorithm 2** Sample augmentation using embedding similarity

**Require:** Training scene graphs $\mathcal{G}^T$, sentence embedding model $f(T)$ for triplet $T$, distance measure $\phi(f(T_a), f(T_b))$, augmentation parameter $\epsilon$, top-$N_e$ object entities $\mathcal{N}_e$, top-$N_r$ relationships $\mathcal{N}_r$.

```
for each G_i ∈ G^T do
    for each R_j(S_j, O_j) ∈ G_i do
        if S_j ∈ N_e and O_j ∈ N_e then
            if R_j ∈ N_r then
                N^a(S_j, R_j, O_j) ← N^a(S_j, R_j, O_j) + 1
            else if R_j ∉ N_r then
                T_i ← concat(S_j, R_j, O_j)
                dist_min ← min_{T←concat(S_j,R,O_j) ∀R∈N_r} φ(T, T_i)
                T_min ← arg min_{T←concat(S_j,R,O_j) ∀R∈N_r} φ(T, T_i)
                if dist_min < ε then
                    S_j, R_min, O_j ← T_min
                    N^a(S_j, R_min, O_j) ← N^a(S_j, R_min, O_j) + 1
                end if
            end if
        end if
    end for
end for
```

**Ensure:** Augmented count $N^a(S, R, O)\ \forall S \in \mathcal{N}_e, \forall R \in \mathcal{N}_r, \forall O \in \mathcal{N}_e s$

## E. Computational complexity

In this section, we compute the computational complexities of MAP inference and constrained optimization.

### E.1. Complexity for MAP inference

For SGCls and SGDet settings, the MAP inference in Eqn. (16) requires inferring $N_s \times N_o \times N_r$ times where $N_s, N_r$, and $N_o$ denote the number of configurations for subject, relationship, and objects. However, the measured probabilities of these three quantities $P_{I,\theta_m}(S), P_{I,\theta_m}(O)$, and $P_{I,\theta_m}(R)$ contain many zeros and hence we choose only top $K_s, K_o$, and $K_r$ values of each triplet measurement. In this way, the computational complexity for top-$K$ triplets in an image becomes $O(K \times K_s \times K_r \times K_o)$.

### E.2. Complexity for constrained optimization

The object updating for each object $E_i$ in Eqn. (26) requires computing probability for $N_e$ object categories and the relationship updating in Eqn. (27) occurs $N_r$ times for each relationship $R_j$. If the number of objects in an image is $N^o$ and we want to infer top-$K$ relationships, the total complexity becomes $O(N^o \times N_e + K \times N_r)$. In the SGCls setting, we choose $K_s = K_o = 3$ for all baseline models. These values for the SGDet setting are provided in Table 7 for all baseline models.

| Baseline Models | IMP [7] | VCTree [6] | MOTIF [8] | Unb-MOTIF [8] | DLFE-MOTIF [2] | BGNN [3] |
|---|---|---|---|---|---|---|
| $K_s, K_o$ | 3 | 3 | 3 | 3 | 1 | 1 |
| $K_r$ | 10 | 10 | 10 | 10 | 50 | 50 |
| $N^o$ | 32 | 32 | 32 | 32 | 80 | 80 |

Table 7. $K_s, K_r$, and $K_o$ for all baseline models in SGDet setting. We choose lower $K_s$ and $K_o$ when $N^o = 80$.

## F. Results and analysis of posterior inference with original and augmented samples

We include a performance with original and augmented samples for all three tasks PredCls, SGCls, and SGDet with Visual Genome and GQA in Table 8. We report both **R@K** and **mR@K**. Baseline SGG models are implemented by the codebase of [5], [2], and [3]. **mR@K** improves in all baseline models and R@K decreases in MOTIF, VCTree, BGNN, and DLFE for VG and in all baseline models for GQA. The decreasing of **R@K** is significantly less in IMP [7], a relatively older method of

SGG. We find this result as promising since the balance between head and tail classes is better achieved with IMP. Inference with original samples improves the **R@K** performance of the debiased predictions from (1) Unb- [5] and (2) DLFE- [2]. We omitted SGCls performance of BGNN [3] since their trained SGCls checkpoint is not released yet. We also include the full comparison with the other de-biasing techniques in Table 9. We achieve better performance in all settings except in SGDet by [2].

| DS | Method | Recall and Mean Recall @K | | | | | |
|---|---|---|---|---|---|---|---|
| | | PredCls | | SGCls | | SGDet | |
| | | R@20/50/100 | mR@20/50/100 | R@ 20/50/100 | mR@20/50/100 | R@20/50/100 | mR@20/50/100 |
| VG | IMP$^{\diamond}$ [7] | 54.9/ 61.6/ 63.6 | 9.2/ 11.5/ 12.4 | 32.9/ 36.1/ 37.1 | 4.9/ 5.7/ 6.0 | 21.0/ 28.0/ 31.3 | 3.3/ 4.9/ 5.8 |
| | Inf-IMP (org) | 55.2/ 62.5/ 64.8 | 15.4/ 20.5/ 22.7 | 33.7/ 37.3/ 38.5 | 7.3/ 9.3/ 10.2 | 20.6/ 27.3/ 30.6 | 4.8/ 7.6/ 9.2 |
| | Inf-IMP (aug) | 53.2/ 59.9/ 62.0 | 18.6/ 25.1/ 28.3 | 32.6/ 36.0/ 37.1 | 9.7/ 12.6/ 14.1 | 20.1/ 26.5/ 29.5 | 5.3/ 8.6/ 10.7 |
| | MOTIF$^{\diamond}$ [8] | 48.9/ 59.6/ 64.0 | 8.4/ 12.9/ 15.5 | 31.2/ 36.5/ 38.5 | 5.5/ 7.7/ 8.8 | 20.7/ 26.9/ 30.5 | 3.9 / 5.6/ 6.7 |
| | Inf-MOTIF (org) | 46.4/ 56.4/ 60.4 | 13.0/ 20.1/ 24.4 | 29.9/ 34.8/ 36.7 | 8.5/ 11.9/ 13.9 | 19.7/ 25.6/ 29.1 | 5.8/ 8.1/ 10.0 |
| | Inf-MOTIF (aug) | 42.5/ 51.5/ 55.1 | 15.7/ 24.7/ 30.7 | 27.7/ 32.2/ 33.8 | 10.2/ 14.5/ 17.4 | 18.6/ 24.0/ 27.1 | 6.6/ 9.4/ 11.7 |
| | VCTree$^{\diamond}$ [6] | 59.1/ 65.5/ 67.2 | 12.0/ 15.4/ 16.6 | 40.4/ 44.2/ 45.1 | 7.4/ 9.2/ 9.8 | 24.0/ 29.9/ 32.6 | 4.7/ 6.2/ 7.0 |
| | Inf-VCTree (org) | 56.6/ 62.5/ 64.1 | 17.7/ 22.7/ 24.8 | 39.3/ 42.9/ 43.8 | 10.7/ 13.5/ 14.6 | 23.4/ 29.1/ 31.6 | 6.3/ 8.4/ 9.5 |
| | Inf-VCTree (aug) | 54.0/ 59.5/ 61.0 | 21.1/ 28.1/ 30.7 | 37.4/ 40.7/ 41.6 | 13.6/ 17.3/ 19.4 | 22.3/ 27.7/ 30.1 | 7.6/ 10.4/ 11.9 |
| | Unb-MOTIF$^{\diamond}$ [5] | 33.4/ 45.9/ 51.2 | 17.9/ 24.8/ 28.7 | 20.5/ 26.3/ 28.8 | 9.8/ 13.2/ 15.1 | 11.8/ 16.3/ 19.5 | 6.4/ 8.7/ 10.5 |
| | Inf-Unb-MOTIF (org) | 34.2/ 47.2/ 52.8 | 18.0/ 25.6/ 30.4 | 21.1/ 27.4/ 30.2 | 10.0/ 13.9/ 16.3 | 12.1/ 17.0/ 20.7 | 6.5/ 9.1/ 11.1 |
| | Inf-Unb-MOTIF (aug) | 31.5/ 42.4/ 46.8 | 19.2/ 28.6/ 35.7 | 19.0/ 24.1/ 26.3 | 10.7/ 15.9/ 18.9 | 10.9/ 15.1/ 18.0 | 6.6/ 9.6/ 11.9 |
| | DLFE-MOTIF$^{\blacklozenge}$ [2] | 45.7/ 51.6/ 53.3 | 22.0/ 26.9/ 28.8 | 25.4/ 28.8/ 29.7 | 12.8/ 15.6/ 16.4 | 18.2/ 24.2/ 28.0 | 8.0/ 10.6/ 12.6 |
| | Inf-DLFE-MOTIF (org) | 49.4/ 55.7/ 57.5 | 25.4/ 31.4/ 33.9 | 27.8/ 31.2/ 32.2 | 14.3/ 17.5/ 18.4 | 20.1/ 26.5/ 30.3 | 9.6/ 12.7/ 14.9 |
| | Inf-DLFE-MOTIF (aug) | 38.0/ 43.3/ 44.8 | 28.5/ 35.3/ 38.2 | 21.4/ 24.3/ 25.1 | 16.3/ 19.7 /20.7 | 15.5/ 20.6/ 23.8 | 10.6/ 14.1/ 16.8 |
| | BGNN$^{\blacklozenge}$ [3] | 50.4/ 58.2/ 60.4 | 24.9/ 29.5/ 31.8 | -/ -/ - | -/ - /- | 23.1/ 30.3/ 35.0 | 7.4/ 10.4/ 12.3 |
| | Inf-BGNN (org) | 50.2/ 57.6/ 59.8 | 25.4/ 30.3/ 33.1 | -/ -/ - | -/ - /- | 22.1/ 28.9/ 33.4 | 8.5/ 12.0/ 14.5 |
| | Inf-BGNN (aug) | 48.2/ 55.4/ 57.5 | 26.2/ 32.2/ 34.3 | -/ -/ - | -/ - /- | 20.0/ 26.2/ 30.1 | 9.4/ 13.2/ 16.1 |
| GQA | IMP$^{\diamond}$ [7] | 57.0/ 61.9/ 63.7 | 11.2/ 13.0/ 13.7 | 32.3/ 34.3/ 34.8 | 6.6/ 7.5/ 7.80 | 20.8/ 25.4/ 27.4 | 4.2/ 5.8/ 6.6 |
| | Inf-IMP (org) | 56.2/ 62.0/ 64.2 | 29.1/ 35.9/ 38.3 | 31.1/ 33.3/ 34.1 | 16.3/ 19.1/ 20.3 | 19.2/ 23.5/ 25.6 | 8.6/ 12.3/ 14.4 |
| | Inf-IMP (aug) | 56.0/ 61.9/ 64.0 | 28.5/ 35.1/ 37.5 | 31.0/ 33.2/ 34.0 | 16.2/ 19.1/ 20.3 | 19.2/ 23.5/ 25.6 | 8.5/ 12.2/ 14.1 |
| | MOTIF$^{\diamond}$ [8] | 64.0/ 68.3/ 69.7 | 17.5/ 20.7/ 21.6 | 33.2/ 34.9/ 35.4 | 9.8/ 10.9/ 11.3 | 23.9/ 27.8/ 29.4 | 5.8/ 7.4/ 8.3 |
| | Inf-MOTIF (org) | 59.0/ 62.9/ 64.1 | 31.9/ 37.8/ 39.8 | 30.3/ 31.9/ 32.4 | 16.8/ 19.0/ 19.9 | 21.9/ 25.5/ 26.9 | 11.6/ 14.4/ 15.9 |
| | Inf-MOTIF (aug) | 59.0/ 63.0/ 64.2 | 32.0/ 37.9/ 40.1 | 30.2/ 31.8/ 32.3 | 16.8/ 19.1/ 20.0 | 21.9/ 25.5/ 26.9 | 11.7/ 14.3/ 15.8 |
| | VCTree$^{\diamond}$ [6] | 64.4/ 68.8/ 70.1 | 18.8/ 22.1/ 23.0 | 33.2/ 35.0/ 35.6 | 9.2/ 10.6/ 11.0 | 23.2/ 27.2/ 28.8 | 5.5/ 7.0/ 7.8 |
| | Inf-VCTree (org) | 59.0/ 62.8/ 64.1 | 33.7/ 39.1/ 41.3 | 30.5/ 32.3/ 32.9 | 16.6/ 19.1/ 19.9 | 21.3/ 25.1/ 26.5 | 11.0/ 13.6/ 15.1 |
| | Inf-VCTree (aug) | 59.0/ 62.8/ 64.1 | 34.1/ 39.4/ 41.6 | 30.5/ 32.2/ 32.8 | 16.6/ 19.2 /20.0 | 21.3/ 25.0/ 26.4 | 11.0/ 13.6/ 15.1 |
| | Unb-MOTIF$^{\diamond}$ [5] | 43.2/ 51.9/ 55.9 | 19.4/ 27.8/ 32.3 | 21.6/ 26.1/ 28.1 | 10.4/ 14.1/ 16.3 | 13.5/ 18.2/ 21.6 | 7.7/ 10.8/ 12.9 |
| | Inf-Unb-MOTIF (org) | 41.4/ 50.1/ 53.9 | 23.1/ 34.9/ 41.3 | 20.6/ 24.9/ 26.9 | 11.9/ 17.3/ 20.7 | 12.5/ 16.8/ 19.9 | 8.7/ 12.4/ 14.8 |
| | Inf-Unb-MOTIF (aug) | 41.4/ 49.9/ 53.6 | 22.9/ 34.5/ 40.8 | 20.5/ 24.8/ 26.7 | 11.9/ 17.2/ 20.6 | 12.5/ 16.8/ 19.9 | 8.7/ 12.4/ 14.8 |

Table 8. Recall@K and Mean Recall@K results of inference with the prefix 'Inf-'. Here, (org) denotes that BN is learned using original training samples and (aug) denotes that BN is learned using augmented samples. $\diamond$ – released by [5], $\blacklozenge$ – released by respective authors. Inference with original samples improves the minority classes and augmentation achieves more improvement over those classes.

## F.1. Qualitative improvement

In Figure 10, a qualitative representation is shown where we see the tail relationships such as 'eating' and 'growing on' are improved with BN inference for different baseline models.

## F.2. Statistical significance test

We partition the full testing dataset into 264 folds with each fold having 100 images. For each fold, we compute the mean recall@100 in PredCls setting for baseline model [6] and for our proposed model. We perform a one-sided Wilcoxon rank sum test between these two groups of performance measures and obtain $p < 0.05$ which denotes that the performance increase in mean recall of our proposed method is statistically significant.

## F.3. Visualization of relationship replacement

In Figure 9, we visualize the statistics of relationship replacement where we observe that through posterior inference the most frequent class 'on' is being replaced by 'sitting on' in $0.7\%$ cases and by 'standing on' in $1.7\%$ cases. Similarly, 'above' is being replaced by 'over' in $35\%$ cases and 'has' is being replaced by 'with' in $8.8\%$ cases.

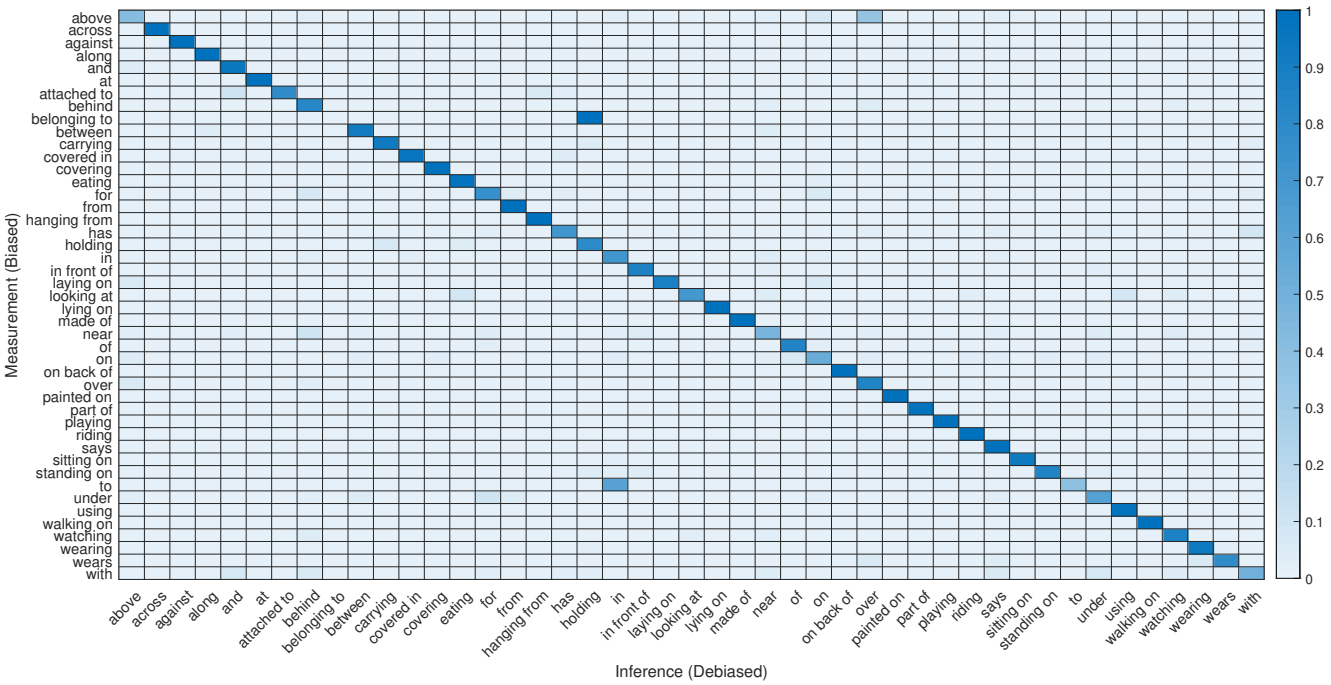



Figure 9. Relationship replacement of measurement by posterior inference. *head* relationships such as 'on', 'has', and 'near' are being replaced by *tail* ones such as 'sitting on', 'with', and 'walking on'.

| **Method** | Recall and Mean Recall @K | | | | | |
|---|---|---|---|---|---|---|
| | PredCls | | SGCls | | SGDet | |
| | R@50/100 | mR@50/100 | R@50/100 | mR@50/100 | R@50/100 | mR@50/100 |
| VCTree [6] | 65.5/ 67.2 | 15.4/ 16.6 | 44.2/ 45.1 | 9.2/ 9.8 | 29.9/ 32.6 | 6.2/ 7.0 |
| Unb-VCTree [5] | 47.2/ 51.6 | 25.4/ 28.7 | 25.4/ 27.9 | 12.2/ 14.0 | 19.4/ 23.2 | 9.3/ 11.1 |
| DLFE-VCTree [2] | 51.8/ 53.5 | 25.3/ 27.1 | 28.0/ 28.9 | 18.2 /19.0 | 22.6/ 26.2 | 11.7/ 13.6 |
| Inf-VCTree (Ours) | 59.5/ 61.0 | 28.1/ 30.7 | 40.7/ 41.6 | 17.3/ 19.4 | 27.7/ 30.1 | 10.4/ 11.9 |

Table 9. Comparison with other de-biasing methods in all settings. Our method achieves significantly higher recall compared to the other debiased methods in PredCls and SGCls without any re-training.

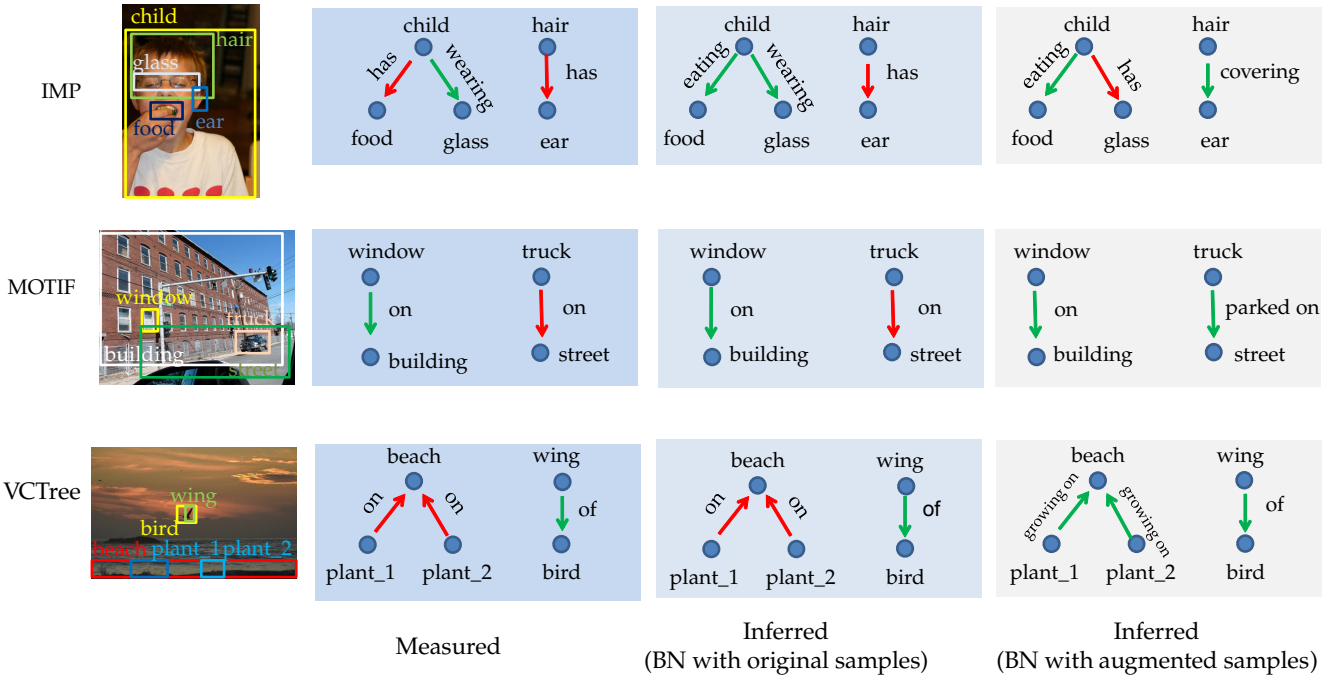



Figure 10. Improvement of tail relationships for IMP [7], MOTIF [8], and VCTree [6]. We visualize the improvement of inference with BN learned from both original and augmented samples. This is a PredCls setting where we know the object locations and classes. The red arrow indicates incorrect relationship and the green arrow indicates correct relationships

## References


[1] Hei Chan and Adnan Darwiche. On the revision of probabilistic beliefs using uncertain evidence. *Artificial Intelligence*, 163(1):67–90, 2005. 1

[2] Meng-Jiun Chiou, Henghui Ding, Hanshu Yan, Changhu Wang, Roger Zimmermann, and Jiashi Feng. Recovering the unbiased scene graphs from the biased ones. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1581–1590, 2021. 5, 6, 7

[3] Rongjie Li, Songyang Zhang, Bo Wan, and Xuming He. Bipartite graph network with adaptive message passing for unbiased scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11109–11119, 2021. 5, 6

[4] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014. 1

[5] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3716–3725, 2020. 5, 6, 7

[6] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6619–6628, 2019. 3, 5, 6, 7, 8

[7] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5419, 2017. 2, 5, 6, 8

[8] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. *arXiv:1711.06640*, 2017. 2, 5, 6, 8