Counterfactual Critic Multi-Agent Training for Scene Graph Generation — Supplementary Document —

This supplementary document is organized as follows:

- Section 1 provides the details of some simplified functions in Agent Communication and Visual Relationship Detection.
- Section 2 provides the detailed proof of the CMAT convergence, which guarantees that the proposed CMAT method can converge to a locally optimal policy.
- Section 3 provides the detailed derivation of Eq. (6), *i.e.*, $\nabla_\theta J \approx \sum_{i=1}^{n} \nabla_\theta \log \boldsymbol{p}_i^T(v_i^T|h_i^T;\theta)Q(H^T,V^T)$.
- Section 4 shows more qualitative results of CMAT compared with the strong baseline MOTIFS [4] in SGDet setting.

1. Details of Some Simplified Functions

We demonstrate the details of some omitted functions in Eq. (1), Eq. (2), Eq. (3) and Eq. (4).

1.1. F_s and F_e in `Extract` Module

$$\begin{aligned} \boldsymbol{h}_i^t &= \text{LSTM}(\boldsymbol{h}_i^{t-1}, [\boldsymbol{x}_i^t, \boldsymbol{e}_i^{t-1}]), \\ \boldsymbol{s}_i^t &= \boldsymbol{s}_i^{t-1} + \boldsymbol{W}_h \boldsymbol{h}_i^t, \\ v_i^t &\sim \boldsymbol{p}_i^t = \text{softmax}(\boldsymbol{s}_i^t), \\ \boldsymbol{e}_i^t &= \textstyle\sum_{\tilde{v}} \boldsymbol{p}_i^t(\tilde{v})\mathbf{E}[\tilde{v}], \end{aligned} \tag{11}$$

where $\boldsymbol{h}_i^t \in \mathbb{R}^h$ is the hidden state of LSTM, $\boldsymbol{x}_i^t \in \mathbb{R}^d$ is the time-step input feature and $\boldsymbol{s}_i^t \in \mathbb{R}^{|\mathcal{C}|}$ is the object class confidence. $\mathbf{E}[\tilde{v}] \in \mathbb{R}^e$ is the embedding of class label $\tilde{v} \in \mathcal{C}$ and $\boldsymbol{e}_i^t \in \mathbb{R}^e$ is the soft-weighted embedding of class label based on probabilities $\boldsymbol{p}_i^t$. $\boldsymbol{W}_h \in \mathbb{R}^{h\times|\mathcal{C}|}$ is a learnable matrix. and $[,]$ is concatenate operation.

1.2. F_{m*} in `Message` Module

$$\boldsymbol{m}_j^t = \boldsymbol{W}_u \boldsymbol{h}_j^t, \ \boldsymbol{m}_{ij}^t = \boldsymbol{W}_p \boldsymbol{h}_{ij}^t \tag{12}$$

where $\boldsymbol{m}_j^t \in \mathbb{R}^h$ and $\boldsymbol{m}_{ij}^t \in \mathbb{R}^h$ is the unary message and pairwise message, respectively. $\boldsymbol{h}_{ij}^t \in \boldsymbol{R}^d$ is the pairwsie feature between agent i and agent j. $\boldsymbol{W}_u \in \mathbb{R}^{h\times h}$, $\boldsymbol{W}_p \in \mathbb{R}^{d\times h}$ are learnable mapping matrices.

1.3. F_{att*} and F_{u*} in `Update` Module

$$\begin{aligned} u_j^t &= \boldsymbol{w}_u[\boldsymbol{h}_i^t, \boldsymbol{h}_j^t], \ \alpha_j^t = \exp(u_j^t)/\textstyle\sum_k \exp(u_k^t), \\ u_{ij}^t &= \boldsymbol{w}_p[\boldsymbol{h}_i^t, \boldsymbol{h}_{ij}^t], \ \alpha_{ij}^t = \exp(u_{ij}^t)/\textstyle\sum_k \exp(u_{ik}^t) \\ \boldsymbol{x}_i^{t+1} &= \boldsymbol{W}_x(\text{ReLU}(\boldsymbol{h}_i^t + \textstyle\sum_j \alpha_j^t \boldsymbol{m}_j^t + \sum_j \alpha_{ij}^t \boldsymbol{m}_{ij}^t)) \\ \boldsymbol{h}_{ij}^{t+1} &= \text{ReLU}(\boldsymbol{h}_{ij}^t + \boldsymbol{W}_s \boldsymbol{h}_i^{t+1} + \boldsymbol{W}_e \boldsymbol{h}_j^{t+1}) \end{aligned} \tag{13}$$

where α_j^t and α_{ij}^t are attention weights to fuse different messages, $\boldsymbol{w}_u \in \mathbb{R}^{2h}$, $\boldsymbol{w}_p \in \mathbb{R}^{h+d}$, $\boldsymbol{W}_x \in \mathbb{R}^{h\times d}$, $\boldsymbol{W}_s \in \mathbb{R}^{h\times d}$, and $\boldsymbol{W}_e \in \mathbb{R}^{h\times d}$ are learnable mapping matrices.

1.4. F_r in Visual Relationship Detection

$$\begin{aligned} \boldsymbol{z}_i &= \boldsymbol{W}_o[\boldsymbol{h}_i^T, \mathbf{E}[v_i^T]], \ \boldsymbol{z}_j = \boldsymbol{W}_o[\boldsymbol{h}_j^T, \mathbf{E}[v_j^T]], \\ \boldsymbol{p}_{ij} &= \text{softmax}([\boldsymbol{z}_i, \boldsymbol{z}_j] \odot \boldsymbol{W}_r \boldsymbol{z}_{ij} + \boldsymbol{w}_{v_i^T, v_j^T}), \\ r_{ij} &= \arg\max_{r\in\mathcal{R}} \boldsymbol{p}_{ij}(r), \end{aligned} \tag{14}$$

where $\boldsymbol{W}_o \in \mathbb{R}^{(h+e)\times z}$, $\boldsymbol{W}_r \in \mathbb{R}^{z\times 2z}$ are transformation matrices, $\boldsymbol{z}_{ij} \in \mathbb{R}^z$ is the predicated visual feature between agent i and j, $\odot$ is a fusing function [1], and $\boldsymbol{w}_{v_i^T, v_j^T} \in \mathbb{R}^{|\mathcal{C}|}$ is the bias vector specific to head and tail labels as in [4].

Predicate Visual Features z_{ij}. For the predicate visual features, we used RoIAlign to pool the union box of subject and object, and resized the union box feature to $7\times7\times512$. Following [4, 1], we used a $14\times14\times2$ binary feature map to model the geometric spatial position of subject and object, with one channel per box. We applied two convolutional layers on this binary feature map and obtained a new $7\times7\times512$ spatial position feature map. We added this position feature map with the previous resized union box feature, and applied two fully-connected layers to obtain the final predicate visual feature.

[1] Different functions get comparable performance. In our experiments, we follow [5]: $\boldsymbol{x} \odot \boldsymbol{y} = \text{ReLU}(\boldsymbol{W}_x\boldsymbol{x} + \boldsymbol{W}_y\boldsymbol{y}) - (\boldsymbol{W}_x\boldsymbol{x} - \boldsymbol{W}_y\boldsymbol{y})^2$.

2. Proof of the Convergence of CMAT

Proof. We denote π_i as the policy of agent i, *i.e.*,$\pi_i = \boldsymbol{p}_i^T$ and $\boldsymbol{\pi}$ as the joint policy of all agents, *i.e.*, $\boldsymbol{\pi} = \{\boldsymbol{p}_1^T, ..., \boldsymbol{p}_n^T\}$. Then, the expected gradient of CMAT is given by (cf. Eq. (11)):

$$\begin{aligned} \nabla_\theta J &= \mathbb{E}_{\boldsymbol{\pi}} \left[\sum_{i=1}^{n} \nabla_\theta \log \pi_i(v_i^T) A^i(H^T, V^T) \right], \\ &= \mathbb{E}_{\boldsymbol{\pi}} \left[\sum_{i=1}^{n} \nabla_\theta \log \pi_i(v_i^T) (R(H^T, V^T) - b(H^T, V_{-i}^T)) \right]. \end{aligned} \tag{15}$$

where the expection $\mathbb{E}_{\boldsymbol{\pi}}$ is with respect to the state-action distribution induced by the joint policy $\boldsymbol{\pi}$, $b(H^T, V_{-i}^T)$ is the counterfactual baseline in CMAT model, *i.e.*, $b(H^T, V_{-i}^T) = \sum \boldsymbol{p}_i^T(\tilde{v}_i^T) R(H^T, (V_{-i}^T, \tilde{v}_i^T))$.

First, consider the expected contribution of this counterfactual baseline $b(H^T, V_{-i}^T)$,

$$\nabla_\theta J_b = \mathbb{E}_{\boldsymbol{\pi}} \left[\sum_{i=1}^{n} \nabla_\theta \log \pi_i(v_i^T) b(H^T, V_{-i}^T) \right]. \tag{16}$$

Let $d^{\boldsymbol{\pi}}(s)$ be the discounted ergodic state distribution as defined by [3]:

$$\begin{aligned} \nabla_\theta J_b =& \sum_s d^{\boldsymbol{\pi}}(s) \sum_{i=1}^{n} \sum_{V_{-i}^T} \boldsymbol{\pi}(V_{-i}^T) \sum_{v_i^T} \pi_i(v_i^T) \nabla_\theta \log \pi_i(v_i^T) b(H^T, V_{-i}^T) \\ =& \sum_s d^{\boldsymbol{\pi}}(s) \sum_{i=1}^{n} \sum_{V_{-i}^T} \boldsymbol{\pi}(V_{-i}^T) \sum_{v_i^T} \nabla_\theta \pi_i(v_i^T) b(H^T, V_{-i}^T) \\ =& \sum_s d^{\boldsymbol{\pi}}(s) \sum_{i=1}^{n} \sum_{V_{-i}^T} \boldsymbol{\pi}(V_{-i}^T) b(H^T, V_{-i}^T) \nabla_\theta 1 \\ =& 0 \end{aligned} \tag{17}$$

Thus, this counterfactual baseline does not change the expected gradient. The reminder of the expected policy gradient is given by:

$$\nabla_\theta J = \mathbb{E}_{\boldsymbol{\pi}} \left[\sum_{i=1}^{n} \nabla_\theta \log \pi_i(v_i^T) R(H^T, V^T) \right], \tag{18}$$

$$= \mathbb{E}_{\boldsymbol{\pi}} \left[\nabla_\theta \log \prod_{i=1}^{n} \pi_i(v_i^T) R(H^T, V^T) \right]. \tag{19}$$

Writing the joint policy into a product of the independent policies:

$$\boldsymbol{\pi}(V^T) = \prod_{i=1}^{n} \pi_i(v_i^T), \tag{20}$$

we have the standard single-agent policy gradient:

$$\nabla_\theta J = \mathbb{E}_{\boldsymbol{\pi}} \left[\nabla_\theta \log \boldsymbol{\pi}(V^T) R(H^T, V^T) \right]. \tag{21}$$

Konda *et al*. [2] proved that this gradient converges to a local maximum of the expected return J, given that: 1) the policy $\boldsymbol{\pi}$ is differentiable, 2) the update timescales for $\boldsymbol{\pi}$ are sufficiently slow. Meanwhile, the parameterization of the policy (*i.e.*, the single-agent joint-action learner is decomposed into independent policies) is immaterial to convergence, as long as it remains differentiable. □

3. Derivation of Eq. (6)

Based on the policy gradient theorem we provide the detailed derivation of Eq. (6) as follows. We denote the action sequence for agent i as $\hat{A}_i = \{\hat{a}_i^1, \hat{a}_i^2, ..., \hat{a}_i^T\}$, and value function $V_\theta(\hat{A}_i)$ as the expected future reward of sequence $\hat{A}_i$. Then the gradient of agent i is:

$$
\begin{aligned}
\nabla_\theta J_i &= \frac{dV(\hat{A}_i)}{d\theta} = \frac{d}{d\theta}\mathbb{E}_{\hat{A}_i\sim\pi_i^t(\hat{A}_i)}R(\hat{A}_i) \\
&= \sum_{\hat{A}_i}\frac{d}{d\theta}\left[\pi_i^t(\hat{a}_i^1)\pi_i^t(\hat{a}_i^2|\hat{a}_i^1)..\pi_i^t(\hat{a}_i^T|\hat{a}_i^1..\hat{a}_i^{T-1})\right]R(\hat{A}_i) \\
&= \sum_{t=1}^{T}\sum_{\hat{A}_i}\pi_i^t(\hat{A}_i^{1..t-1})\frac{d\pi_i^t(\hat{a}_i^t|\hat{A}_i^{1..t-1})}{d\theta}\pi_i^t(\hat{A}_i^{t+1..T}|\hat{A}_i^{1..t})R(\hat{A}_i) \\
&= \sum_{t=1}^{T}\sum_{\hat{A}_i^{1..t}}\pi_i^t(\hat{A}_i^{1..t-1})\frac{\pi_i^t(\hat{a}_i^t|\hat{A}_i^{1..t-1})}{d\theta}\sum_{\hat{A}_i^{t+1..T}}\pi_i^t(\hat{A}_i^{t+1..T}|\hat{A}_i^{1..t})\sum_{\tau=1}^{T}r_i^\tau(\hat{a}_i^\tau;\hat{A}_i^{1..\tau-1}) \\
&= \sum_{t=1}^{T}\sum_{\hat{A}_i^{1..t}}\pi_i^t(\hat{A}_i^{1..t-1})\frac{\pi_i^t(\hat{a}_i^t|\hat{A}_i^{1..t-1})}{d\theta}\left[r^t(\hat{a}_i^t;\hat{A}_i^{1..t-1}) + \sum_{\hat{A}_i^{t+1..T}}\pi_i^t(\hat{Y}_i^{t+1..T}|\hat{Y}_i^{1..t})\sum_{\tau=t+1}^{T}r_i^\tau(\hat{a}_i^\tau;\hat{A}_i^{1..\tau-1})\right] \\
&= \sum_{t=1}^{T}\mathbb{E}_{\hat{A}_i^{1..t-1}\sim\pi_i^t(\hat{A}_i^{1..t-1})}\sum_{a^t\in\mathcal{A}}\frac{d\pi_i^t(a^t|\hat{A}_i^{1..t-1})}{d\theta}Q(a^t;\hat{A}_i^{1..t-1}) \\
&= \mathbb{E}_{\hat{A}_i\sim\pi_i^t(\hat{A}_i)}\sum_{t=1}^{T}\sum_{a^t\in\mathcal{A}}\frac{\pi_i^t(a^t|\hat{A}_i^{1..t-1})}{d\theta}Q(a^t;\hat{A}_i^{1..t-1})
\end{aligned}
\tag{22}
$$

Further, the gradient for agent i can be simplified as:

$$
\begin{aligned}
\nabla_\theta J_i &= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a_i^t\in\mathcal{A}}\nabla_\theta\pi_i^t(a_i^t)Q(s_i^t,a_i^t)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a_i^t\in\mathcal{A}}\pi_i^t(a_i^t)\nabla_\theta\log\pi_i^t(a_i^t)Q(s_i^t,a_i^t)\right] \\
&\approx \sum_{t=1}^{T}\nabla_\theta\log\pi_i^t(a_i^t)Q(s_i^t,a_i^t)
\end{aligned}
\tag{23}
$$

Therefore, for the time step t, the gradient for agent i is $\nabla_\theta\log\pi_i^t(a_i^t)Q(s_i^t,a_i^t)$. For multi-agent in a cooperative environment, the state-action function Q should estimate the reward based on the set of all agent state and actions, *i.e.*, $Q(S^t,A^t)$. Then, the gradient for all agents is:

$$
\nabla_\theta J \approx \sum_{i=1}^{n}\nabla_\theta J_i = \sum_{i=1}^{n}\nabla_\theta\log\pi_i^t(a_i^t|s_i^t)Q(S^t,A^t). \tag{24}
$$

In CMAT, we samples actions after T-round agent communication, and the action for agent i is $\boldsymbol{v}_i^T$, the policy function is $\boldsymbol{p}_i^T$, and the state of agent is $\boldsymbol{h}^t$, *i.e.*, $S^t = H^t$, $A^t = V^t$. Therefore, the gradient for the cooperative multi-agent in CMAT is:

$$
\nabla_\theta J \approx \sum_{i=1}^{n}\nabla_\theta\log\boldsymbol{p}_i^T(v_i^T|h_i^T;\theta)Q(H^T,V^T) \tag{25}
$$

4. More Qualitative Results

Figure 7 and 8 show more qualitative results of CMAT and MOTIFS in SGDet setting. From the rows where CMAT is better than MOTIFS, we can see that CMAT rarely mistakes at the important hub nodes such aas the “surfboard” or “laptop”.

This is because CMAT directly optimizes the graph-coherent objective. However, the rows show that the mistakes made by CMAT always come from the imcomplete anntation of CMAT can detect more false positive (the blue color) objects and relationship than MOTIFS. Since the evaluation metric (i.e., Recall@K) is based on the ranking of labeled triplet confidence, thus, detecting more reasonable false positive results with high confidence can worsen the performance.

References

[1] Bo Dai, Yuqi Zhang, and Dahua Lin. Detecting visual relationships with deep relational networks. In *CVPR*, 2017. 1

[2] Vijay R Konda and John N Tsitsiklis. Actor-critic algorithms. In *NeurIPS*, 2000. 2

[3] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *NeurIPS*, 2000. 2

[4] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *CVPR*, 2018. 1

[5] Yan Zhang, Jonathon Hare, and Adam Prügel-Bennett. Learning to count objects in natural images for visual question answering. In *ICLR*, 2018. 1

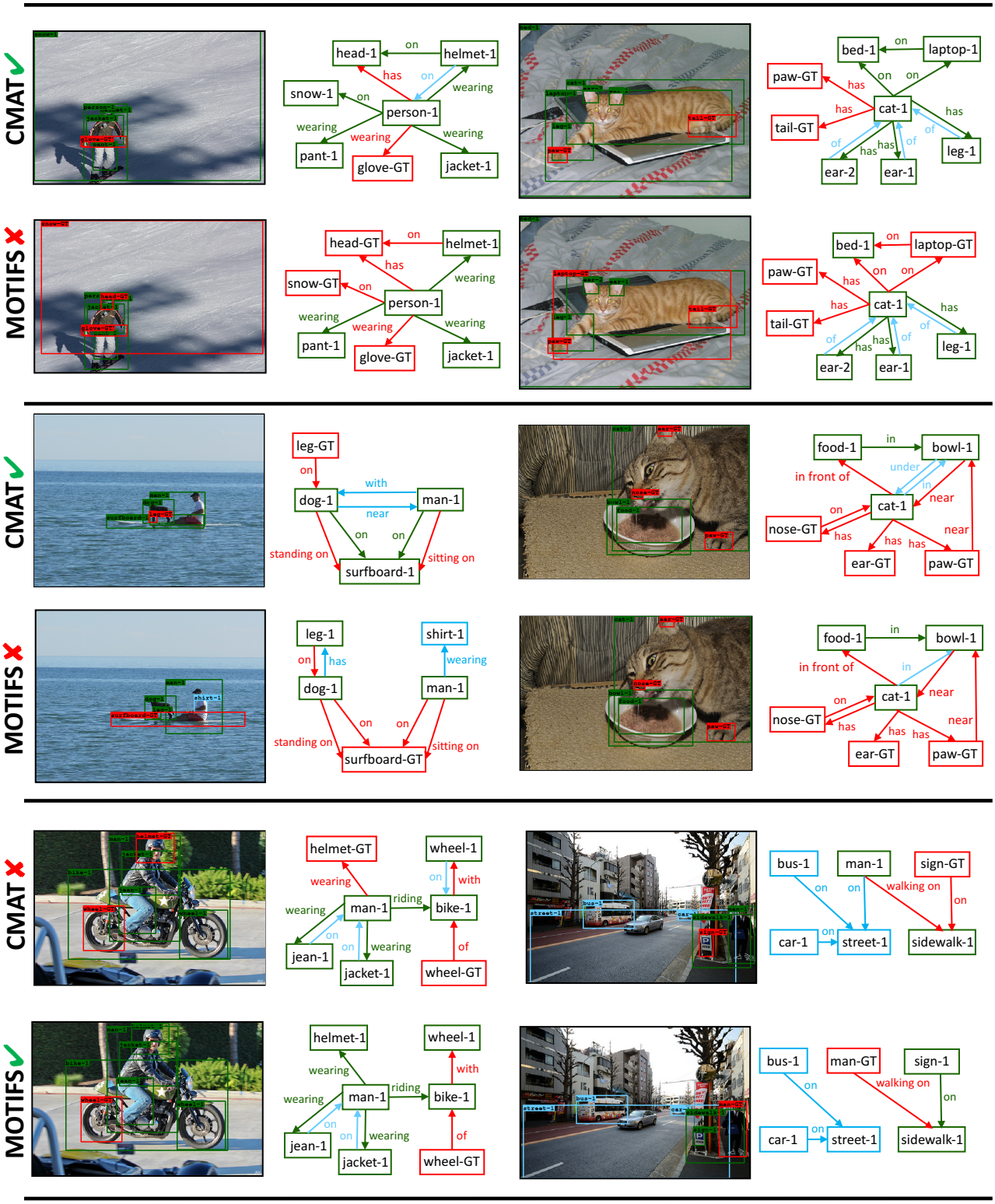

Figure 7: More qualitative results showing comparisons between CMAT and MOTIFS in the SGDet setting. Green boxes are detected boxes with IoU large 0.5 with the ground truth, blue boxes are detected but not labeled, red boxes are ground-truth with no match. Green edges are true positive predicted by each model at the R@20 setting, red edges are false negatives, and blue edges are false positives. Only detected boxes overlapped with ground-truth are shown.

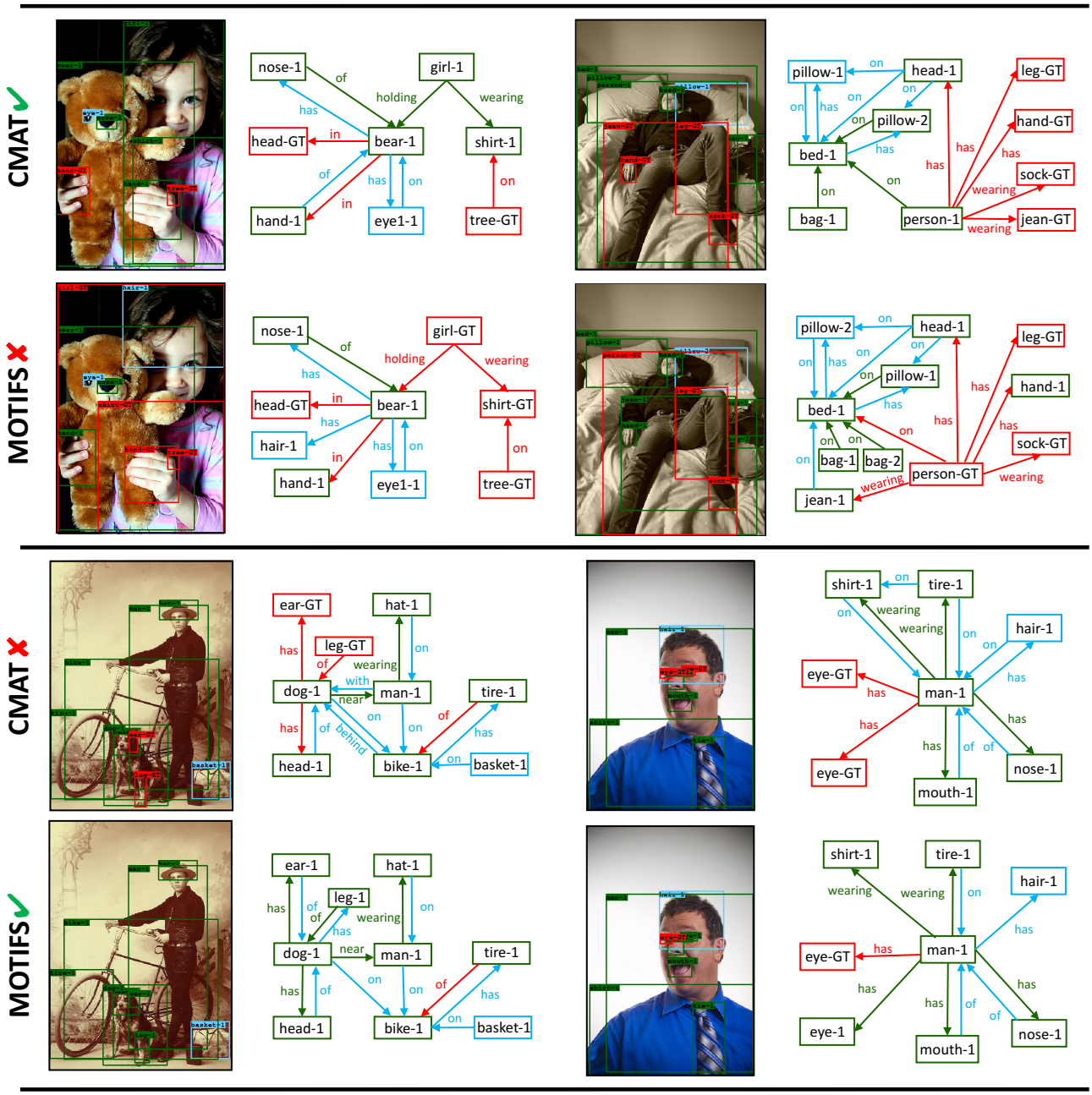

Figure 8: More qualitative results showing comparisons between CMAT and MOTIFS in the SGDet setting continued from Figure 7.