

# Sketching Image Gist: Human-Mimetic Hierarchical Scene Graph Generation

Wenbin Wang<sup>1,2[0000-0002-4394-0145]</sup>, Ruiping Wang<sup>1,2[0000-0003-1830-2595]</sup>,  
Shiguang Shan<sup>1,2[0000-0002-8348-392X]</sup>, and Xilin Chen<sup>1,2[0000-0003-3024-4404]</sup>

<sup>1</sup> Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, 100190, China

<sup>2</sup> University of Chinese Academy of Sciences, Beijing, 100049, China

[wenbin.wang@vipl.ict.ac.cn](mailto:wenbin.wang@vipl.ict.ac.cn), {wangruiping, sgshan, xlchen}@ict.ac.cn

In this supplementary material, we include the following sections:

- Section 1: We provide some cases to explain why the so-called image gist or key relations are not equal to visually salient objects or where humans gaze.
- Section 2: Detailed introduction about transforming our HetH and HetH-RRM models for image captioning and re-implementation of the GCN-LSTM [7] model.
- Section 3: More details about VG-KR construction and additional statistics of our VG200 and VG-KR datasets. Besides, the implementation details and hyperparameters settings are provided.
- Section 4: The statistical significance of the results from our method.
- Section 5: We give an example to show the advantages of HET.
- Section 6: The exploration and findings on VG-KR, which inspire us that both visual saliency and size of an object are helpful for estimating relation importance.
- Section 7: More qualitative results of our method.

## 1 Detailed Explanation about Motivation

As illustrated in the main paper, it's notable that the visually salient objects are related but not completely equal to objects involved in image gist. According to findings in [1], objects referred in a description (i.e., objects that humans think important and should form the major events/image gist) are almost visually salient and reveal where humans gaze, but what humans fixate (i.e., visually salient objects) are not always what they want to convey at first. In Figure 1, we provide some examples to show that this is a common phenomenon. E.g., the *red clothes*, the *Spring Festival couplets*, and the *black doors of the washing machines* (mentioned from left to right), are visually salient due to their high contrast to the context or center position. However, some of them do not form the major events. For example, in the 2<sup>nd</sup> image, the first glance description would be “There stands a house on the side of the road”. Then humans may be interested in the eyecatching *Spring Festival couplets*.

Besides, we are inspired by these observations. There naturally exists a hierarchical structure about humans' perception preference. Objects with relatively

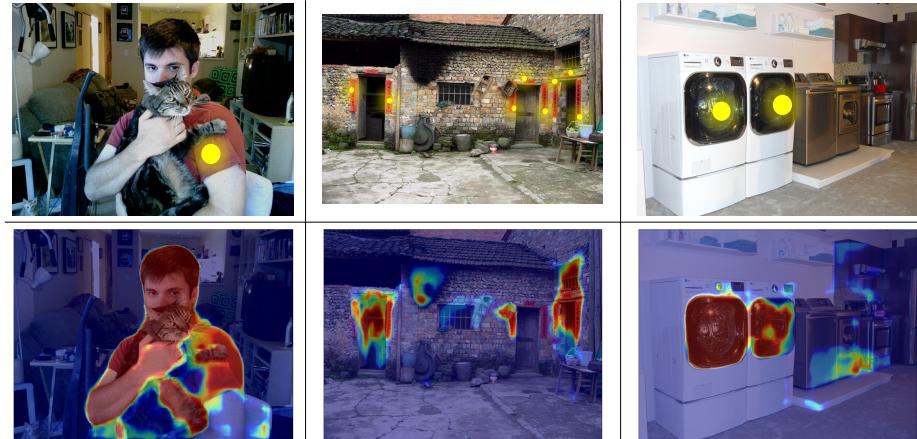


Fig. 1: Visually salient objects do not always form the major events in the images and are not always what humans want to convey at first from the images. The yellow points in each image denote some visually salient objects. The saliency maps in the second row are obtained from [2].

large size which fulfill the scene generally form the major events. It supports us to construct HET with the method introduced in the main paper. We aim at constructing HET whose levels reflect the perception priority level rather than the object saliency. The experiments show that our method for constructing HET has achieved this goal.

## 2 Implementation Details for Image Captioning

As the source codes of GCN-LSTM [7] have not been released by the submission deadline, we re-implement it. In its original version, a simple two-layer MLP classifier is applied to predict the pairwise relationship, which acts as the frontend scene graph detector. For a fair comparison, we replace this detector with our HetH. To transform our HetH/HetH-RRM for image captioning task, we add a sentence decoder which is modified from LSTM backend of GCN-LSTM. The GCN-LSTM model conducts graph convolution on the scene graph and injects all relation-aware **region**-level features into a two-layer LSTM with attention mechanism. Different from GCN-LSTM, we intend to inject the relation features rather than region-level features, considering that the relationships which convey the events in the image are more helpful for description generation. In Figure 2, we show a brief diagram to illustrate our implementation of GCN-LSTM, and demonstrate the implementation scheme of our sentence decoder for image captioning.

Specifically, we obtain a set of visual relationship representations  $\{\mathbf{f}_m^R\}_{m=1}^M$  ( $\mathbf{f}_m^R \in \mathbb{R}_f^D$ ,  $D_f = 4,096$ ) after relation context decoding (see Figure 2(d) in the

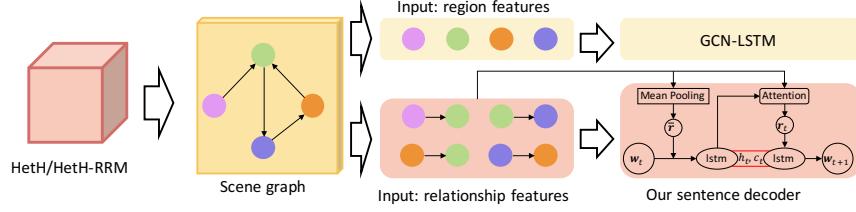


Fig. 2: Our implementation of GCN-LSTM, and the the implementation scheme of our sentence decoder.

main paper). We concatenate them with the word embeddings of their subjects, objects, and predicates, denoted by  $\mathbf{w}_m^s \in \mathbb{R}^{D_w}$ ,  $\mathbf{w}_m^o \in \mathbb{R}^{D_w}$ , and  $\mathbf{w}_m^p \in \mathbb{R}^{D_w}$  ( $D_w = 300$ ), and obtain  $\{\mathbf{r}_m\}_{m=1}^M$  ( $\mathbf{r}_m \in \mathbb{R}^{D_r}$ ,  $D_r = 4,996$ ):

$$\mathbf{r}_m = [\mathbf{f}_m^{\mathcal{R}}; \mathbf{w}_m^s; \mathbf{w}_m^o; \mathbf{w}_m^p]. \quad (1)$$

The sentence decoder is a two-layer LSTM. It's noted that two layers in this decoder share one hidden state  $\mathbf{h} \in \mathbb{R}^{D_h}$  and cell state  $\mathbf{c} \in \mathbb{R}^{D_h}$ . At each time step  $t$ , the first layer collects the maximum contextual information by concatenating the input word embedding  $\mathbf{w}_t \in \mathbb{R}^{D_w}$  and the mean-pooled visual relationship feature  $\bar{\mathbf{r}} = \frac{1}{M} \sum_{m=1}^M \mathbf{r}_m$ . The updating procedure is as

$$\mathbf{h}_t^1, \mathbf{c}_t^1 = f_1([\mathbf{w}_t; \bar{\mathbf{r}}])_{|\mathbf{h}_{t-1}^2, \mathbf{c}_{t-1}^2}, \quad (2)$$

where  $f_1$  is the updating function within the first-layer unit,  $|\mathbf{h}_{t-1}^2, \mathbf{c}_{t-1}^2$  denotes that the internal hidden state and cell state is the ones that updated by the second-layer unit from the previous timestep. Then we compute a normalized attention distribution over all the relationship features

$$a_{t,m} = \mathbf{W}_a [\tanh (\mathbf{W}_f \mathbf{r}_m + \mathbf{W}_h \mathbf{h}_t^1)], \quad \lambda_t = \text{softmax}(\mathbf{a}_t), \quad (3)$$

where  $a_{t,m}$  is the  $m$ -th element of  $\mathbf{a}_t$ ,  $\mathbf{W}_a \in \mathbb{R}^{1 \times D_a}$ ,  $\mathbf{W}_f \in \mathbb{R}^{D_a \times D_r}$ ,  $\mathbf{W}_h \in \mathbb{R}^{D_a \times D_h}$  are transformation matrices. Specifically, both the dimension of the hidden layer  $D_a$  for measuring the attention distribution and the dimension of the hidden layer  $D_h$  in LSTM are set as 512.  $\lambda_t \in \mathbb{R}^M$  denotes the normalized attention distribution whose  $m$ -th element  $\lambda_{t,m}$  is the attention weight of  $\mathbf{r}_m$ .

The attended relationship feature is computed as  $\mathbf{r}_t = \sum_{m=1}^M \lambda_{t,m} \mathbf{r}_m$ . Then the updating procedure of the second-layer unit is

$$\mathbf{h}_t^2, \mathbf{c}_t^2 = f_2(\mathbf{r}_t)_{|\mathbf{h}_t^1, \mathbf{c}_t^1}, \quad (4)$$

where  $f_2$  is the updating function within the second-layer unit.  $\mathbf{h}_t^2$  is used to predict the next word through a softmax layer.

Table 1: Statistics of VG200, VG-KR, and VG150.

Dataset	Images	Images with Relations	Object Categories	Predicate Categories	Object Instances	Relation Instances	Key Relation Instances	Images with Key Relation Instances
VG200	51,498	46,562	200	80	619,119	442,425	101,312	26,992
VG-KR	26,992	26,992	200	80	360,306	250,755	-	-
VG150	108,073	89,169	150	50	1,145,398	622,705	-	-

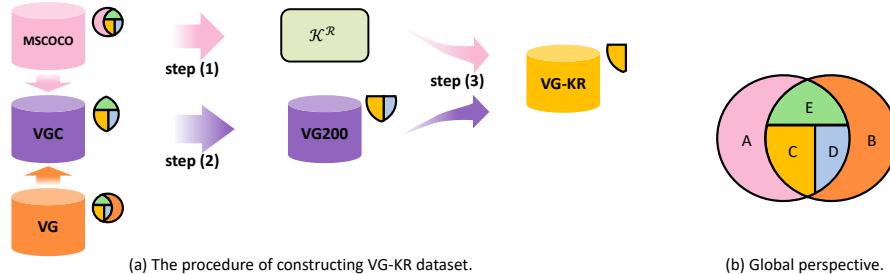


Fig. 3: The procedure of VG-KR dataset construction. The color block shown in the top right of each dataset in (a) demonstrates its components. (b) gives a global perspective of these color blocks. Images in MSCOCO consist of four parts, A, C, D, and E. Images in VG consist of B, C, D, and E. Images in VGC consist of C, D, and E. E denotes images filtered in step (2) which do not contain any relation. D denotes images filtered in step (3) which do not contain any key relation.

### 3 VG-KR Dataset Construction, Statistics, and Experimental Implementation Details

#### 3.1 VG-KR Dataset

We demonstrate the procedure of constructing VG-KR dataset and different image sets involved in this procedure in Figure 3. Concretely, 51,498 images in VG come from MSCOCO, which form the image set  $VGC$ . We conduct three-stage processing on  $VGC$ . (1) Stanford Scene Graph Parser [4] is used to extract relation triplets from captions. They make up the set of key relations, denoted by  $\mathcal{R}^K$ . (2) We next cleanse the raw annotations of  $VGC$  similar to [6], keep the most frequent 150 object categories and 50 predicates, and add another most frequent 50 object categories and 30 predicates in  $\mathcal{R}^K$ , in order to keep as many key relations as possible for the following third step. After dropping images without relations in  $VGC$ , we get a new subset **VG200** (i.e. 200 object categories) which contains 46,562 images. (3) Finally, we associate  $\mathcal{R}^K$  with relation triplets in VG200 by associating their subject and object WordNet synsets [3] respectively. After filtering out the images without key relations in VG200, here comes **VG-KR** which contains 26,992 images. For both VG200 and VG-KR, we split the

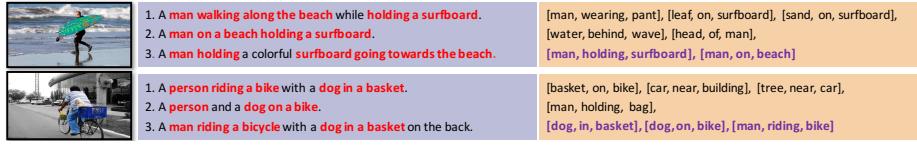


Fig. 4: Examples in VG-KR dataset. Each image is shown with 3 captions and ground truth relations. Purple triplets are key ones while others are secondary.

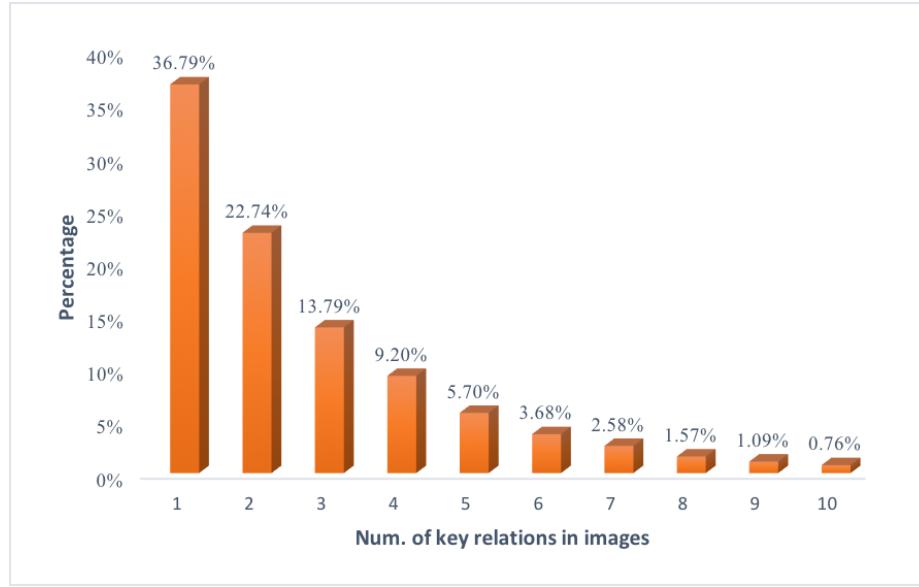


Fig. 5: Distribution of images that contain different numbers of key relations.

training and test set by 7:3 ratio, leading to 32,510/14,052 training/test images in VG200, and 18,720/8,272 training/test images in VG-KR.

We give several examples in Figure 4 and show more detailed statistics and compare with VG150 in Table 1. We can see that VG200 and VG-KR have more categories, as well as object and relation instances per image compared to VG150. Moreover, VG-KR contains indicative annotations of key relations.

In Figure 5, we show the distribution of images that contain different numbers of key relations. More than 90% of images contain less than 5 key relations. It's reasonable because the key relations are obtained by matching the annotated relations with those extracted from captions. The number of relation triplets in captions generally is not very large. After all, a good caption is only requested to describe the major contents instead of the less important details.

Given each predicate, we explore the distribution of its role, i.e., whether it belongs to a key relation or not. The result is shown in Figure 6. The predicates with large probability to be key ones, such as *throwing*, *brushing*, and *sniffing*, are usually verbs containing rich semantics. They are image-specific and when

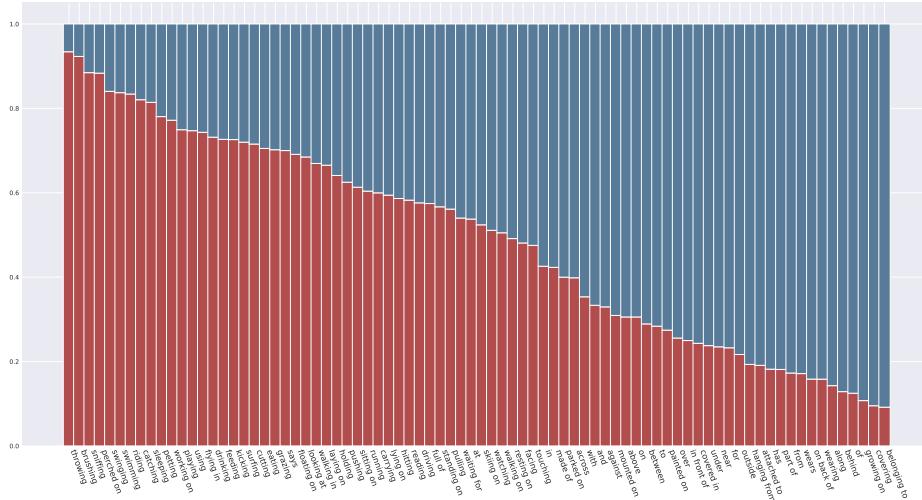


Fig. 6: Distribution of the roles of a given predicate. The red bars stand for the probability of being key relations while the blue bars denote the probability of being the secondary ones.

we see these predicates, a scene can be roughly imagined. While predicates like *belonging to*, *of*, and *behind*, which carry little information, are less likely to make up the key relations.

### 3.2 Settings and Implementation Details

The dimension of hidden states and cells in both Hybrid-LSTM and RRM is 512. The sizes of  $\mathbf{W}_1^{(r)}$  and  $\mathbf{W}_2^{(r)}$  in Eq.(11) in the main paper are  $256 \times 512$  and  $1 \times 256$  respectively. The GloVe embedding vectors we use are of 200 dimensions.

When training on the VG dataset, we follow previous works [8,5] to extract the first 5,000 images of the training split and treat them as the validation split. The results reported on VG150, VG200, and VG-KR are obtained by firstly selecting the best model on validation split and then evaluating it on test split. As for the experiments on VRD, we report the results of the last epoch evaluated on test split without model selection (The hyperparameters settings are the same as those of experiments on VG).

We pre-train object detectors on VRD, VG150, and VG200 respectively and freeze the learned parameters. To train the whole model end-to-end, we use an SGD optimizer with a learning rate of 0.001 and the batch size is 10. When computing the ranking loss for RRM, we randomly sample 512 pairs of key triplets and secondary triplets. The margin  $\gamma$  is empirically set to 0.5. All the existing methods evaluated on our VG200 or VG-KR datasets are retrained.

The threshold  $T$  for determining a parent node actually has direct influence on the shape of HET. We investigate the performance curve together with the tree depth and width variation trend. As shown in Figure 7, as  $T$  varies from

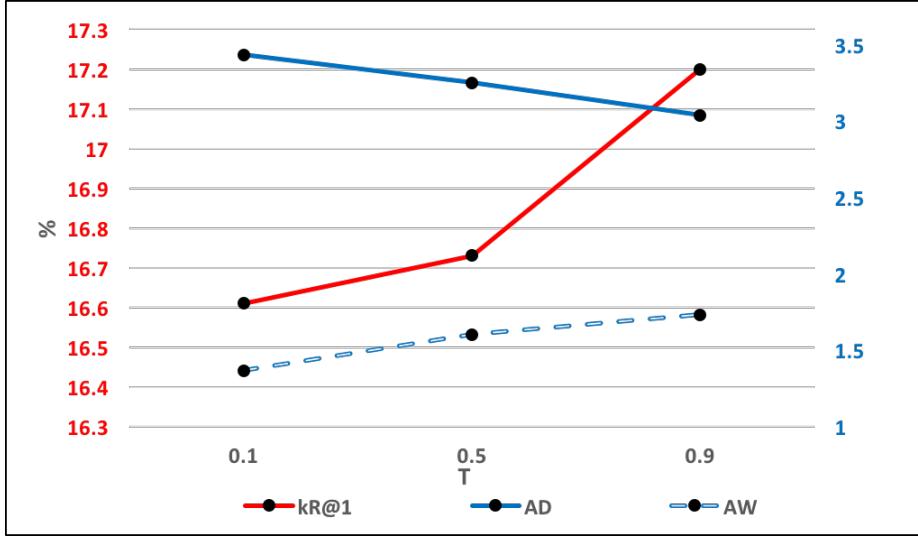


Fig. 7: Effect of threshold  $T$  when constructing HET. AD and AW denote average depth and average width respectively. The red curve stands for the  $kR@1$  performance of HET-RRM, evaluated under PREDCLS protocol with triplet-match rule.

0.1 to 0.9, the “tall thin” tree becomes a “short fat” tree, and the performance is improved. Thus we set  $T$  to 0.9.

As  $T$  becomes larger, the condition that a node can be a parent node, i.e.,  $P_{nm} > T$  (Eq.(1) in the main paper), is more and more difficult to be satisfied. Thus, our algorithm for constructing HET tends to set the root as the parent of a node, which results in a “shorter” and “fatter” tree.

A small  $T$  would lead to considerable wrong hierarchical connections. It’s noted that the hierarchical connections in our HET have much stronger semantics than the associations of siblings. Therefore, a large  $T$  eliminates wrong hierarchical connections as far as possible. Although it means that more entities are set as the child of the root and inappropriate siblings associations increase, proper hierarchical connections still plays a positive role in context encoding.

## 4 Robustness Analyses

We make multiple runs on the HetH under the PREDCLS protocol. The results and statistical significance are shown in Table 2.

Table 2: The results of multiple runs of HetH and the statistical significance. These results are obtained under the PREDCLS protocol.

#RUN	R@20	R@50	R@100
1	33.46	36.59	37.00
2	33.53	36.64	37.04
3	33.93	36.65	37.07
$\mu \pm \sigma$	$33.64 \pm 0.21$	$36.63 \pm 0.03$	$37.04 \pm 0.03$

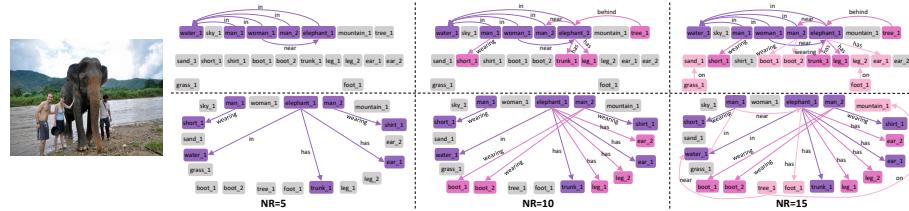


Fig. 8: As the quota of top relations (NR) increases, scene graphs dynamically enlarge. The newly involved entities and relations are shown in a new color. Results in first row and second row are from HetH-RRM and MOTIFS respectively.

## 5 Advantages of HET

As shown in the first row in Figure 8, hierarchical scene graph from our HetH-RRM enlarges in a top-down manner as the quota of top relations increases, while the ordinary scene graph in the second row enlarges itself aimlessly. If we want to limit the scale of a scene graph but keep its ability to sketch image gist as far as possible, it is feasible for our hierarchical scene graph since we just need to cut off some secondary branches of HET, but is difficult to realize in an ordinary scene graph.

## 6 Exploration on VG-KR

We develop the Relation Ranking Module (RRM) to prioritize key relations. We intend to capture humans’ subjective assessment on the importance of relations with some objective indicators. As analyzed in Section 1, visually salient objects engage humans’ gaze and have the potential to form major events. Therefore, visual saliency can be one of the useful indicators. However, it’s easy to lead to misunderstandings when only visual saliency is considered.

To better describe the importance of relation, we borrow the traditional “saliency” concept, and put forward a brand-new concept, *cognitive saliency*, which tries to estimate the importance of a relation from humans’ perspective

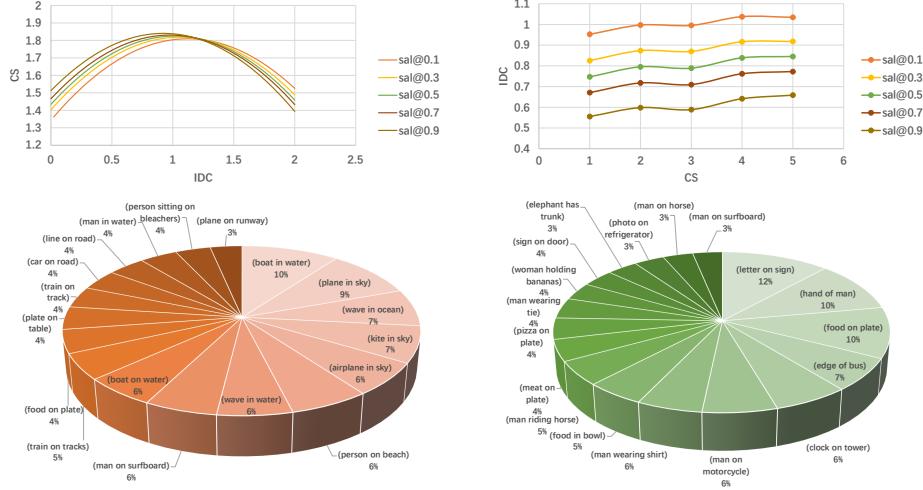


Fig. 9: Curve charts and pie charts for the indicator which is the sum of subject and object visual saliency values. In the curve charts, different curves are drawn under different thresholds  $T_s$ . **Top left:** CS - IDC chart. **Top right:** IDC - CS chart. **Bottom left:** Component analysis for relations which have small IDC values. **Bottom right:** Component analysis for relations which have large IDC values.

as the sensation of importance of relation is very subjective. Considering the measurement of cognitive saliency of a relation triplet, we employ its *times being referred within the five captions* of each image, which can be directly obtained during the construction of our VG-KR dataset. However, this measurement of cognitive saliency is not computable. (i.e., it is grading from humans, but cannot be directly used in computational models.) If we want to make use of the cognitive saliency, we need to find a computable indicator for it. The indicator should be proportional to cognitive saliency, which means that as the cognitive saliency goes up, the same trend should be observed on the indicator, and vice versa.

Intuitively, the first possible indicator is the visual saliency of subject and object in a relation triplet. Specifically, we set the indicator  $\Phi$  as the sum of saliency values of subject  $o^{sub}$  and object  $o^{obj}$ :

$$\mathcal{S}^{sub} = \frac{|\{p|p \in \mathbf{b}^{sub} \wedge \mathcal{S}^p > T_s\}|}{|\{p|p \in \mathbf{b}^{sub}\}|}, \quad (5)$$

$$\mathcal{S}^{obj} = \frac{|\{p|p \in \mathbf{b}^{obj} \wedge \mathcal{S}^p > T_s\}|}{|\{p|p \in \mathbf{b}^{obj}\}|}, \quad (6)$$

$$\Phi = \mathcal{S}^{sub} + \mathcal{S}^{obj}, \quad (7)$$

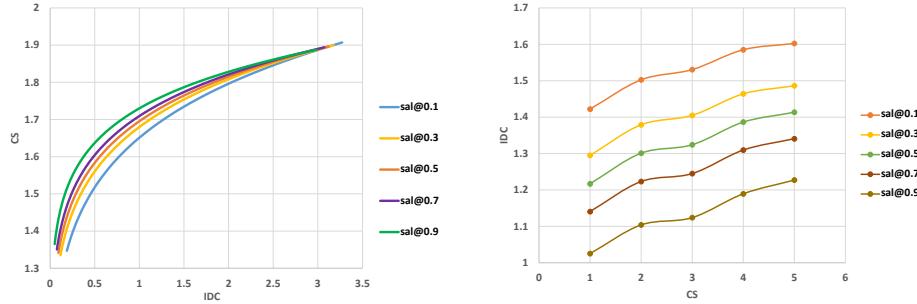


Fig. 10: Curve charts for the indicator which is the sum of subject and object visual saliency values and normalized areas. Different curves are drawn under different thresholds  $T_s$ . **Left:** CS - IDC chart. **Right:** IDC - CS chart.

where  $p$  denotes pixels,  $\mathbf{b}^{sub}$  and  $\mathbf{b}^{obj}$  are bounding boxes of subject and object,  $S^p$  is the saliency value of pixel  $p$ ,  $S^{sub}$  and  $S^{obj}$  are saliency values of subject and object,  $T_s$  is a given threshold.  $|\cdot|$  computes the number of elements in a set. The pixel-wise saliency is computed by one of the state-of-the-art saliency detectors [2].

To draw the Cognitive Saliency (CS, Y-axis) - Indicator (IDC, X-axis) curve chart, we randomly sampled 50,000 key relations from VG-KR with grading from 1 to 5 as their CS values. As IDC values (i.e.,  $\Phi$  in Eq. (7)) are continuous, we sort all the sampled IDC values in ascending order, and divide them into 50 intervals  $[\delta_k, \delta_{k+1}], 0 \leq k \leq 50$ , where  $\delta_0 = IDC_{min}$  and  $\delta_{50} = IDC_{max}$ . In each interval, we draw a point with the mean of the sampled IDC values as X-axis coordinate and the mean of sampled CS values as Y-axis coordinate. When it comes to IDC (Y-axis) - CS (X-axis) curve chart, the sampled relations are grouped by CS values. We compute the mean of IDC values for each group as Y-axis coordinates. These two charts are shown in Figure 9. In each chart, we draw curves under different settings of  $T_s$ , denoted by  $sal@T_s$ . From the IDC - CS chart at the top right of Figure 9, IDC is proportional to CS. However, the CS - IDC chart at the top left of Figure 9 shows that CS is not strictly proportional to IDC, which means that although the computed visual saliency of an object is large, the relations involved in this object are not so important. What results in this phenomenon? We further extract the relations with relatively small IDC values and large IDC values respectively and analyze the ratio of each type of triplet. Concretely, we find the quartering points  $\lambda_1 < \lambda_2 < \lambda_3$  of IDC values, and all the triplets whose IDC values are smaller than  $\lambda_1$  or larger than  $\lambda_3$  are picked out, namely the set  $\Psi$  and  $\Omega$ . The component analysis results of the set  $\Psi$  and  $\Omega$  are shown at the bottom of Figure 9, where the most frequent 18 types of triplets are demonstrated. From the bottom left pie chart, lots of triplets with low IDC and low CS values generally are relations between relatively small objects and the large background entities. However, there are some exceptions,

e.g.,  $\langle man, on, surfboard \rangle$ , and  $\langle train, on, tracks \rangle$ . It's reasonable and we should explore the detailed image contents if we want to further analyze the association between their IDC and CS values. What we should pay attention to is the bottom right pie chart, where we observe that most triplets in  $\Omega$  are relations between an independent object and its components, such as  $\langle hand, of, man \rangle$  and  $\langle edge, of, bus \rangle$ . Actually, these relations are indeed not so image-specific and carry little information. Humans generally overlook them. However, if the saliency of an object is large, saliency of its component will be large, too. It explains the phenomenon when IDC keeps increasing, the CS decreases instead.

In order to further rectify the indicator above, we consider the size of subject and object out of the thinking that the sizes of components or details of a certain entity is relatively small, which can balance the large saliency value. Therefore, we add the normalized size of subject and object into the indicator:

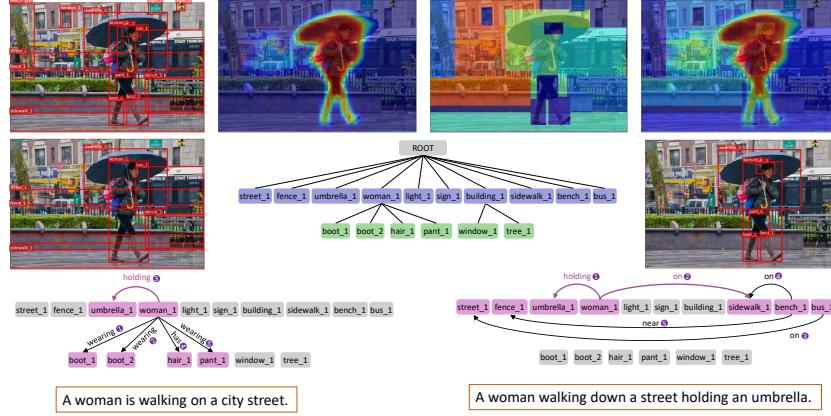
$$\Phi' = \mathcal{S}^{sub} + \mathcal{S}^{obj} + \frac{A(o^{sub})}{A(o_{\mathcal{I}})} + \frac{A(o^{obj})}{A(o_{\mathcal{I}})}, \quad (8)$$

where  $A(\cdot)$  denotes the size function, and  $o_{\mathcal{I}}$  denotes the whole image. Similarly, we draw the IDC - CS and CS - IDC charts in Figure 10. It is shown that this improved indicator is a feasible one, as the CS is strictly proportional to IDC, and vice versa.

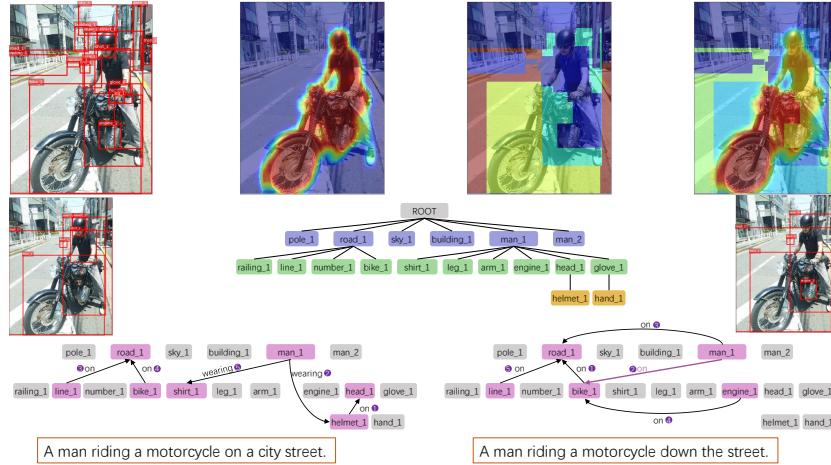
The exploration above inspires us that an indicator which contains both the visual saliency and size of an object may be useful for finding key relations. Therefore, our devised RRM learns to capture humans' subjective assessment on the importance of relations under the guidance of visual saliency and entity size information.

## 7 Additional Qualitative Results

We demonstrate more qualitative results in Figure 11. From all of these examples, it can be seen that our RRM tends to describe relations between entities which are close to the root of HET. These relations describe the global contents and usually are what humans pay the most attention to. As a result, the captions generated from top relations better cover the essential contents. For example, in Figure 11(a), as the top-2 relations from HetH model contain  $\langle woman, wearing, boot\_1/boot\_2 \rangle$ , the generated caption cannot capture the essential content that the *woman* is *holding* an *umbrella*. On the contrary, top-2 relations from HetH-RRM successfully capture this information. In some cases, we observe that although top-2 relations do not contain the essential content, the generated caption can still capture it, e.g., the caption from HetH in Figure 11(b). It is mainly because the region of *man\_1* contains part of the region of *motorcycle\_1*, which provides visual cues for inferring the content that a *man* is *riding* a *motorcycle*.



(a)



(b)

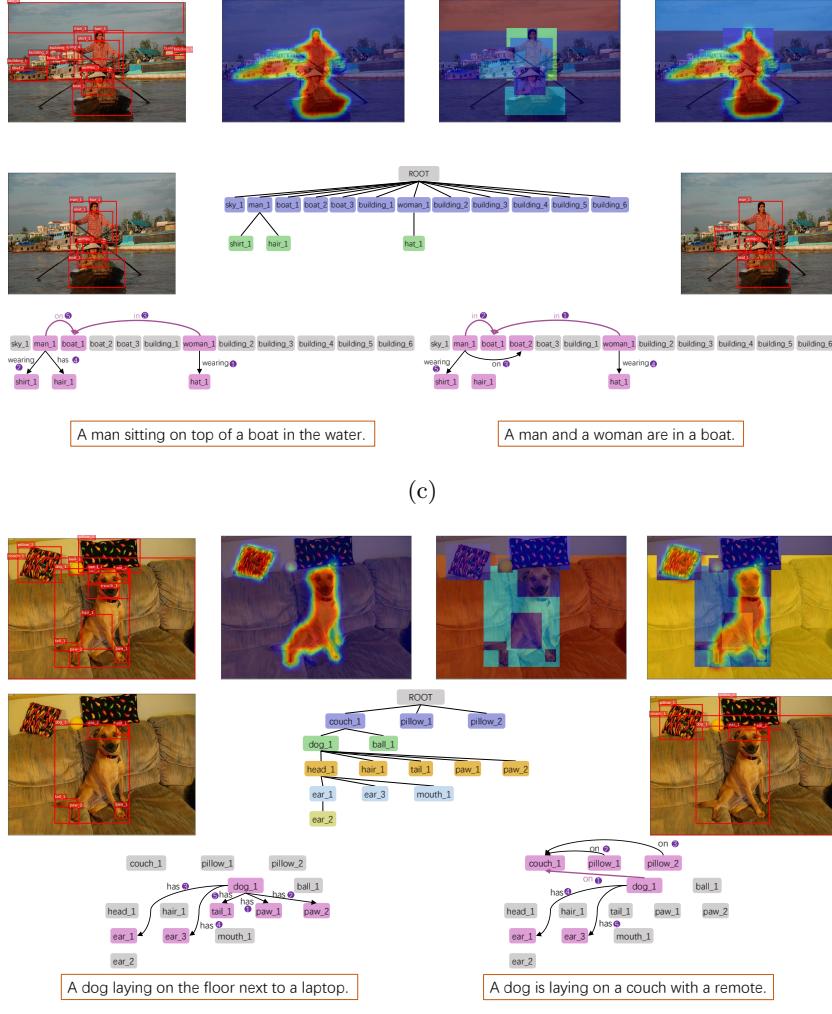


Fig. 11: From top left to bottom right are: bounding boxes of all objects, saliency maps, area maps, mixed maps, bounding boxes of objects involved in top-5 relations from HetH, HET structure, bounding boxes of objects involved in top-5 relations from HetH-RRM model, hierarchical scene graphs from HetH and HetH-RRM model, generated captions using top-2 relations from HetH and HetH-RRM respectively. The purple arrows in scene graphs are key relations matched with ground truth. The purple numeric tags next to the relations are the rankings, and “1” means that the relation gets the highest score.

## References

1. He, S., Tavakoli, H.R., Borji, A., Pugeault, N.: Human attention in image captioning: Dataset and analysis. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 8529–8538 (2019) 1
2. Hou, Q., Cheng, M.M., Hu, X., Borji, A., Tu, Z., Torr, P.: Deeply supervised salient object detection with short connections. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3203–3212 (2017) 2, 10
3. Miller, G.A.: Wordnet: A lexical database for english. *Communication of the ACM* **38**(11), 39–41 (1992) 4
4. Schuster, S., Krishna, R., Chang, A., Fei-Fei, L., Manning, C.D.: Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In: Proceedings of the Fourth Workshop on Vision and Language. pp. 70–80 (2015) 4
5. Tang, K., Zhang, H., Wu, B., Luo, W., Liu, W.: Learning to compose dynamic tree structures for visual contexts. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6619–6628 (2019) 6
6. Xu, D., Zhu, Y., Choy, C.B., Fei-Fei, L.: Scene graph generation by iterative message passing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5410–5419 (2017) 4
7. Yao, T., Pan, Y., Li, Y., Mei, T.: Exploring visual relationship for image captioning. In: Proceedings of European Conference on Computer Vision (ECCV). vol. 11218, pp. 711–727. Springer (2018) 1, 2
8. Zellers, R., Yatskar, M., Thomson, S., Choi, Y.: Neural motifs: Scene graph parsing with global context. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5831–5840 (2018) 6