Conformal Prediction and MLLM aided Uncertainty Quantification in Scene Graph Generation
Supplementary Material

Sayak Nag[1], Udita Ghosh[1], Calvin-Khang Ta[2]*, Sarosij Bose[1], Jiachen Li[1], Amit K. Roy-Chowdhury[1]
[1]University of California, Riverside, USA, [2]Dolby Laboratories, USA
{snag005,ughos002,sbose007,jiachen.li}@ucr.edu, calvin.ta@dolby.com, amitrc@ece.ucr.edu

A. Algorithmic and Mathematical Details

A.1. Algorithm for Conformal Scene Graph Generation

The overall algorithm for our proposed SGG-specific CP is shown in Algorithm 1. The following details are important regarding the algorithm:

- The training and calibration sets are denoted as $\mathcal{D}_{tr} = \{I_j, G_j\}_{j=1}^{m}$ and $\mathcal{D}_{cal} = \{I_j, G_j\}_{j=m+1}^{m+n}$, where I_j is an image and G_j is its ground truth scene graph. Each G is composed of triplets belonging to the object and predicate classes $(\mathcal{Y}_o, \mathcal{Y}_r)$ as described in Sec 3.2. A test image is denoted as I_{n+1}.

- It must be noted that when referring to a sample, X_j, for a prediction set $\hat{\mathcal{C}}(X_j)$, we do not refer to the whole image. Specifically for the object prediction set $\hat{\mathcal{C}}_o(X_i)$, X_j refers to a single RoI describing a detected object. For the predicate prediction set $\hat{\mathcal{C}}_r(X_j)$, the sample X_j refers to a pair of detected objects, as information about the pair including the union box of its two objects is included in the sample. In some cases like in Algorithm 1 we specifically distinguish the samples for the object and predicate classes by denoting them as X_j^o and X_j^r respectively (the meaning remains the same).

- The assumption of exchangeability i.e. $\mathcal{D}_{cal} \cup (I_{n+1}, G_{n+1})$ also implies exchangeability holds for any subsets of D_{cal}, such as the considered partitions $\mathcal{D}^o_{cal,y} \subset \mathcal{D}_{cal}$, and $\mathcal{D}^r_{cal,y} \subset \mathcal{D}_{cal}$.

- Algorithm 1 can sometimes result in null sets for the objects and predicates, however, such an event occurred extremely rarely during our experiments. In such cases, we follow common practice [10] and choose the class with the highest softmax probability value i.e. $\underset{y \in \mathcal{Y}}{max}\ \pi_y$, as the prediction set. As such a singleton prediction set is constructed in such cases.

*Work done while at UCR.

Algorithm 1 Conformal Scene Graph Generation

```
Input: Training Set: $\mathcal{D}_{tr}$, Calibration Set: $\mathcal{D}_{cal}$, Object Miscoverage Rate: $\alpha_o$, Predicate Miscoverage Rate: $\alpha_r$,
  Test Image: $I_{n+1}$.
Output: Object Prediction Set: $\hat{\mathcal{C}}_o(X^o_{n+1})$, Predicate Prediction Set: $\hat{\mathcal{C}}_r(X^r_{n+1})$, Triplet Prediction Set: $\hat{\mathcal{C}}_t(X^r_{n+1})$.

Fit an SGG model, $\phi$, on the training set $\mathcal{D}_{train}$.
Calibration Procedure:
Assume the calibration set $\mathcal{D}_{cal}$ which is comprised of images has the following subsets,
    $\mathcal{D}^o_{cal,y} = \{(X^o_i, Y_i) : Y_i = y^o_i\}, \forall y^o_i \in \mathcal{Y}_o$.
    $\mathcal{D}^r_{cal,y} = \{(X^r_i, Y_i) : Y_i = y^r_i\}, \forall y^r_i \in \mathcal{Y}_r$.
  where, $\mathcal{D}^o_{cal,y}, \mathcal{D}^r_{cal,y} \in \mathcal{D}_{cal}$ define a classwise calibration subset specific to the object and predicate classes in the
  images of $\mathcal{D}_{cal}$, and $X^o_i \in \mathcal{D}^o_{cal,y}$ refers to the object classification specific sample defined by an RoI in the image,
  $X^r_i \in \mathcal{D}^o_{cal,y}$ refers to the predicate classification specific sample defined by a pair of objects [3, 5, 11].
Define a nonconformity function
    $\mathcal{A} : \mathcal{X} \times \mathcal{Y} \to [0, 1]$, $(\hat{f}(X), y) \mapsto 1 - \hat{\pi}_y(X)$
  $\hat{f}$ is any classifier, $\hat{\pi}_y(X)$ is estimated true class probability. Therefore, the complement of $\hat{\pi}_y(X)$ encodes a notion of
  dissimilarity (nonconformity) between the predicted and true class probabilities.
Define list of object and predicate class quantiles $Q_o$ and $Q_r$.
Match pair of detected objects with ground-truth pair of objects using Eq 3.
Begin for each $y^o_i \in \mathcal{Y}_o$ and $y^r_i \in \mathcal{Y}_r$:
Apply $\mathcal{A}$ to $\mathcal{D}^o_{cal,y}$ to obtain a set of scores
    $S_{y^o_i} = \{\mathcal{A}(f_o(X_i), y^o_i)\}_{i=1}^{n_{y^o_i}} = \{s^o_i\}_{i=1}^{n_{y^o_i}}$. where $f_o$ is the object classifier within $\phi$
Apply $\mathcal{A}$ to $\mathcal{D}^r_{cal,y}$ to obtain a set of scores
    $S_{y^r_i} = \{\mathcal{A}(f_r(X_i), y^r_i)\}_{i=1}^{n_{y^r_i}} = \{s^r_i\}_{i=1}^{n_{y^r_i}}$. where $f_r$ is the predicate classifier within $\phi$
Compute a conformal quantiles $\hat{q}_{y^o_i}$ and $\hat{q}_{y^r_i}$, defined as,
    $\hat{q}_{y^o_i} = \lceil (n_{y^o_i} + 1)(1 - \alpha_o)/n_{y^o_i} \rceil$-th empirical quantile of $S_{y^o_i}$.
    $\hat{q}_{y^r_i} = \lceil (n_{y^r_i} + 1)(1 - \alpha_r)/n_{y^r_i} \rceil$-th empirical quantile of $S_{y^r_i}$.
Add object class quantile to the set: $Q_o = Q_o \cup \{\hat{q}_{y^o_i}\}$, and Add predicate class quantile to the set: $Q_r = Q_r \cup \{\hat{q}_{y^r_i}\}$.
End for
End procedure
Conformal Inference Procedure:
For a new test Image $I_{n+1}$ comprised of objects depicted by $(X^o_{n+1}, Y^o_{n+1})$, and predicates depicted by $(X^r_{n+1}, Y^r_{n+1})$,
  valid prediction sets for $X^o_{n+1}$ and $X^r_{n+1}$ are constructed as,
    $\hat{\mathcal{C}}_o(X^o_{n+1}) = \{y^o_k \in \mathcal{Y}_o : \hat{\pi}_{y^o_k} \geq 1 - \hat{q}_{y^o_k}\}$
    $\hat{\mathcal{C}}_r(X^r_{n+1}) = \{y^r_k \in \mathcal{Y}_r : \hat{\pi}_{y^r_k} \geq 1 - \hat{q}_{y^r_k}\}$
  where $\hat{q}_{y^r_k} \in Q_r$ and $\hat{q}_{y^o_k} \in Q_o$. The validity of the sets refers to satisfying class-conditional coverage guarantee (Eq 2)
  with probability $(1 - \alpha_o)$ for each object class, and $(1 - \alpha_r)$ for each predicate class. (Proof: Sadinle et al. [8])
Combinatorially combine $\hat{\mathcal{C}}_o(X^o_{n+1})$ and $\hat{\mathcal{C}}_r(X^r_{n+1})$ to construct a triplet prediction $\hat{\mathcal{C}}_t(X^r_{n+1})$.
End procedure
```

A.2. Nominal Coverage Guarantee of Triplet prediction Sets

Theorem 1. *Given the ground truth class of the k^{th} triplet is denoted as $y^t_k = [y^s_k, y^r_k, y^o_k] \in \mathbb{R}^3$ where $y^s_k, y^o_k \in \mathcal{Y}_o$ and $y^r_k \in \mathcal{Y}_r$, the triplet coverage guarantee is given as $P(y^t_k \in \hat{\mathcal{C}}_t(X^r_{n+1})) = P(Y^o_{n+1} \in \hat{\mathcal{C}}_o(X^o_{n+1}) \mid Y^o_{n+1} = y^o_i) \cdot P(Y^r_{n+1} \in \hat{\mathcal{C}}_o(X^r_{n+1}) \mid Y^r_{n+1} = y^r_m) \; \forall \; y^s_k \in \mathcal{Y}_o, y^o_k \in \mathcal{Y}_o, y^r_k \in \mathcal{Y}_r$.*

Proof. Assume the ground truth of the k^{th} triplet in an image is denoted as $y^t_k = [y^s_k, y^r_k, y^o_k] \in \mathbb{R}^3$ where $y^s_k, y^o_k \in \mathcal{Y}_o$ are the ground-truth classes of the subject-object pair and $y^r_k \in \mathcal{Y}_r$ is the ground-truth class of the predicate. Since $y^s_k, y^o_k \in \mathcal{Y}_o$, for clarity let's denote $y^s_k = y^o_i \mid i \in [1, K_o]$, $y^o_k = y^o_j \mid j \in [1, K_o]$ & $j \neq i$, and $y^r_k = y^r_m \mid m \in [1, K_r]$. Now assuming the prediction sets for the subject and object are given as $\hat{\mathcal{C}}_o(X^o_{n+1})$, $\hat{\mathcal{C}}_o(X^o_{n+2})$, and the prediction set of the predicate is given as $\hat{\mathcal{C}}_r(X^r_{n+1})$, the nominal coverage guarantee of a triplet prediction set is described as,

$$P(y^t_k \in \hat{\mathcal{C}}_t(X^r_{n+1})) = P(Y^o_{n+1} \in \hat{\mathcal{C}}_o(X^o_{n+1}) \wedge Y^o_{n+2} \in \hat{\mathcal{C}}_o(X^o_{n+2}) \wedge Y^r_{n+1} \in \hat{\mathcal{C}}_r(X^r_{n+1}) | Y^o_{n+1} = y^o_i, Y^o_{n+2} = y^o_j, Y^r_{n+1} = y^r_m) \quad \text{(i)}$$

We observe that the guarantees of the object and predicate prediction are controlled by distinct conformal procedures on the calibration data and, as such, are conditionally independent. Therefore,

$$P(y^t_k \in \hat{\mathcal{C}}_t(X^r_{n+1})) = P(Y^o_{n+1} \in \hat{\mathcal{C}}_o(X^o_{n+1}) \wedge Y^o_{n+2} \in \hat{\mathcal{C}}_o(X^o_{n+2}) | Y^o_{n+1} = y^o_i, Y^o_{n+2} = y^o_j) \cdot P(Y^r_{n+1} \in \hat{\mathcal{C}}_o(X^r_{n+1}) | Y^r_{n+1} = y^r_m) \quad \text{(ii)}$$

Additionally, there is no separate subject and object detection, as all objects in an image are detected once and then combinatorially combined to form subject-object pairs. Therefore, the class-conditional coverage guarantee of the subject is contained in the class-conditional coverage guarantee of the object. Hence,

$$P(Y^o_{n+1} \in \hat{\mathcal{C}}_o(X^o_{n+1}) \wedge Y^o_{n+2} \in \hat{\mathcal{C}}_o(X^o_{n+2}) | Y^o_{n+1} = y^o_i, Y^o_{n+2} = y^o_j) = P(Y^o_{n+1} \in \hat{\mathcal{C}}_o(X^o_{n+1}) | Y^o_{n+1} = y^o_i) \quad \text{(iii)}$$

$$\implies P(y^t_k \in \hat{\mathcal{C}}_t(X^r_{n+1})) = P(Y^o_{n+1} \in \hat{\mathcal{C}}_o(X^o_{n+1}) \mid Y^o_{n+1} = y^o_i) \cdot P(Y^r_{n+1} \in \hat{\mathcal{C}}_o(X^r_{n+1}) \mid Y^r_{n+1} = y^r_m) \quad \text{(iv)}$$

□

Corollary 1. *Following Theorem 1,* $P(y^t_k \in \hat{\mathcal{C}}_t(X^r_{n+1})) \geq (1 - \alpha_o)(1 - \alpha_r)$, $\forall\, y^s_k \in \mathcal{Y}_o, y^o_k \in \mathcal{Y}_o, y^r_k \in \mathcal{Y}_r$.

Proof. This follows trivially from Theorem 1. Since

$$P(Y^o_{n+1} \in \hat{\mathcal{C}}_o(X^o_{n+1}) \mid Y^o_{n+1} = y^o_i) \geq (1 - \alpha_o) \quad and \quad P(Y^r_{n+1} \in \hat{\mathcal{C}}_o(X^r_{n+1}) \mid Y^r_{n+1} = y^r_m) \geq (1 - \alpha_r) \quad \text{(v)}$$

$$\implies P(y^t_k \in \hat{\mathcal{C}}_t(X^r_{n+1})) \geq (1 - \alpha_o) \cdot (1 - \alpha_r) \quad \text{(vi)}$$

□

It must be pointed out that the nominal coverage guarantee in Eq vi is the intended coverage goal, based on which calibration is conducted (under assumptions of exchangeability). However, in practice, the empirical coverage may not always reach coverage guarantees, owing to the high predictive uncertainty of the underlying model as well as, unquantified/subtle distribution shifts between the train/calibration and test data [1, 7, 8, 10]. We observe this in our empirical results (Table 1). Additionally, from an empirical standpoint, the results in Table 1, of the main paper, validate Eq iv in the sense that the empirical coverage of the triplet prediction set is approximately close to the product of the empirical coverages of the object and predicate prediction sets.

B. Additional Implementation Details

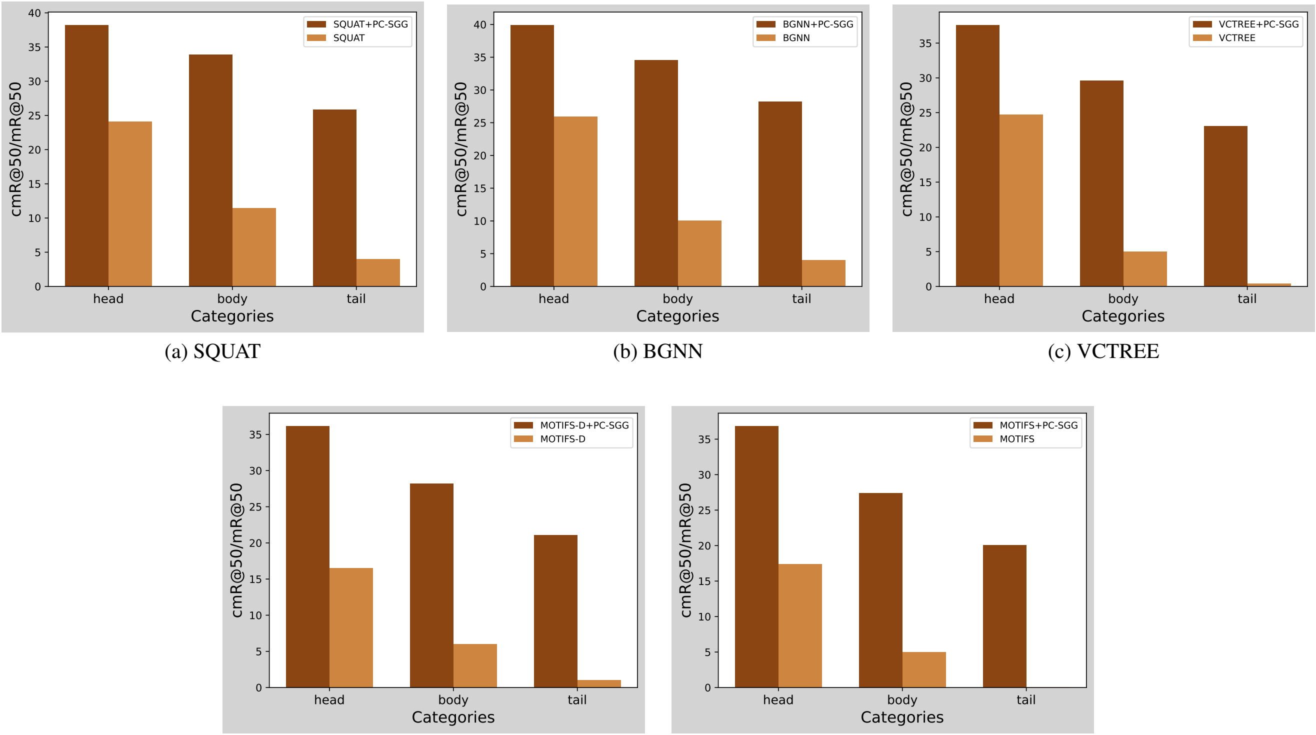
B.1. Computation of Conformal Prediction Metrics

End-to-end SGG or SGDET entails the localization and classification of all objects in a scene along with classifying their pairwise predicates. As such, when computing the CP metrics on the inference data, we need to match the predicted bounding boxes with the ground-truth ones in an image. We do so by following the same pairwise greedy matching strategy shown in Eq 3. Finally, the CP metrics are computed over the matched predictions only as is standard practice in CP-based localization studies [6, 10].

B.2. Adaptation of Scene Graph Generation Metrics for Prediction Sets

For SGG the Recall@K (R@K) and mean-Recall@K (mR@K) are standard evaluation metrics [9]. Formally if K predicted triplets $\{\hat{t}_i\}^K_{i=1}$, are matched to a ground-truth triplet defined by the class y_i, then R@K is computed as,

$$R@K = \frac{1}{N} \sum_{i=1}^{N} \left(\bigvee_{j=1}^{K} \mathbb{1}[y_i = \hat{t}_j] \right) \quad \text{(vii)}$$



(a) SQUAT

(b) BGNN

(c) VCTREE

(d) MOTIFS-D

(e) MOTIFS

Figure I. Improvement of recall-hit rate across head, body, and tail categories of VG150 [4], with the incorporation of PC-SGG. The darker shade is the cmr@50 value and the lighter one is the mR@50 value.

where, $\bigvee$ is the logical OR operation, $\mathbb{1}[\cdot]$ is the indicator function, N is the total number of ground-truth triplets, and $\hat{t}_j, y_j \in \mathbb{R}^3$. However, R@$K$ is not designed for prediction sets and so to accommodate triplet prediction sets we propose coverage-Recall@K (cR@K) which is computed as follows,

$$cR@K = \frac{1}{N}\sum_{i=1}^{N}\left(\bigvee_{j=1}^{K}\mathbb{1}[y_i \in \hat{\mathcal{C}}_{t,j}(X_i^r)]\right) \tag{viii}$$

where $\hat{\mathcal{C}}_{t,j}(X_i^r)$ is the triplet prediction set of the j^{th} predicted triplet. Therefore, the only difference between R@K and cR@K is that the *equality* ($=$) operation is replaced with the *belongs to* ($\in$) operation. As such cR@K is equivalent to R@K when predictions are not in the form of a single triplet but in the form of prediction sets. The coverage-mean-Recall@K (cmR@K) metric is designed similarly and is equivalent to the mR@K metric.

C. Additional Empirical Results

Table I. **Impact of number of options in the MCQA prompt.** Results are shown for the BGNN model. The number of options here refers to the number of prediction set entries. The ‘no valid option’ choice is not counted.

Number of MCQA Options	$Cov_T \uparrow$	$AvgSize \downarrow$
Original (w/o MLLM post-processing)	80.45	971.69
3	**80.45**	539.72
5	**80.45**	464.11
10	69.38	**291.47**

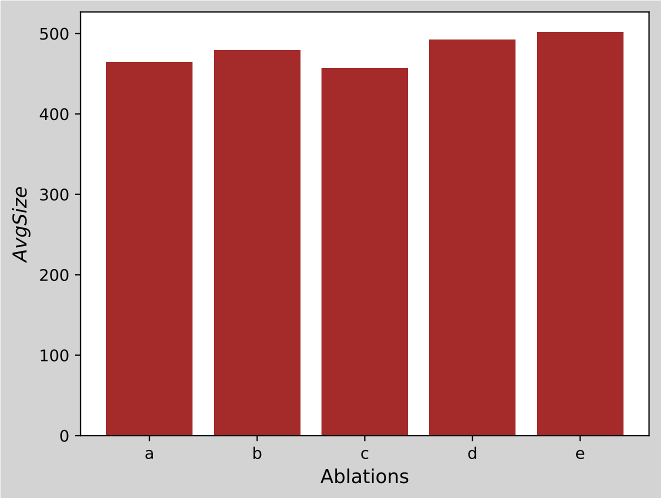


Figure II. **Impact of different prompting strategies on $AvgSize$ of triplet prediction sets.** (a)Proposed prompting design, (b) No Image Cropping, (c) No System Prompt, (d) No Example prompt, (e) No System and Example prompt. All results are obtained for the BGNN model.

C.1. Per Class Performance Improvement

Incorporating PC-SGG with any existing SGG method significantly improves the recall-hit rate for every class in the scene graph dataset. this is evident from Fig 4 in the main paper where we show the massive performance improvement over some of the tail classes of VG150 [4] for the BGNN method. In Fig I we further show the aggregate improvement of recall-hit rate across the HEAD, BODY, and TAIL classes of VG150 when PC-SGG is added to any of the SGG methods used in this paper. Specifically for MOTIFS and VCTREE, which had negligible or zero mR@50 values for the TAIL classes, incorporating PC-SGG is significantly beneficial as the generated triplet prediction sets cover most of the TAIL classes.

C.2. More Analysis

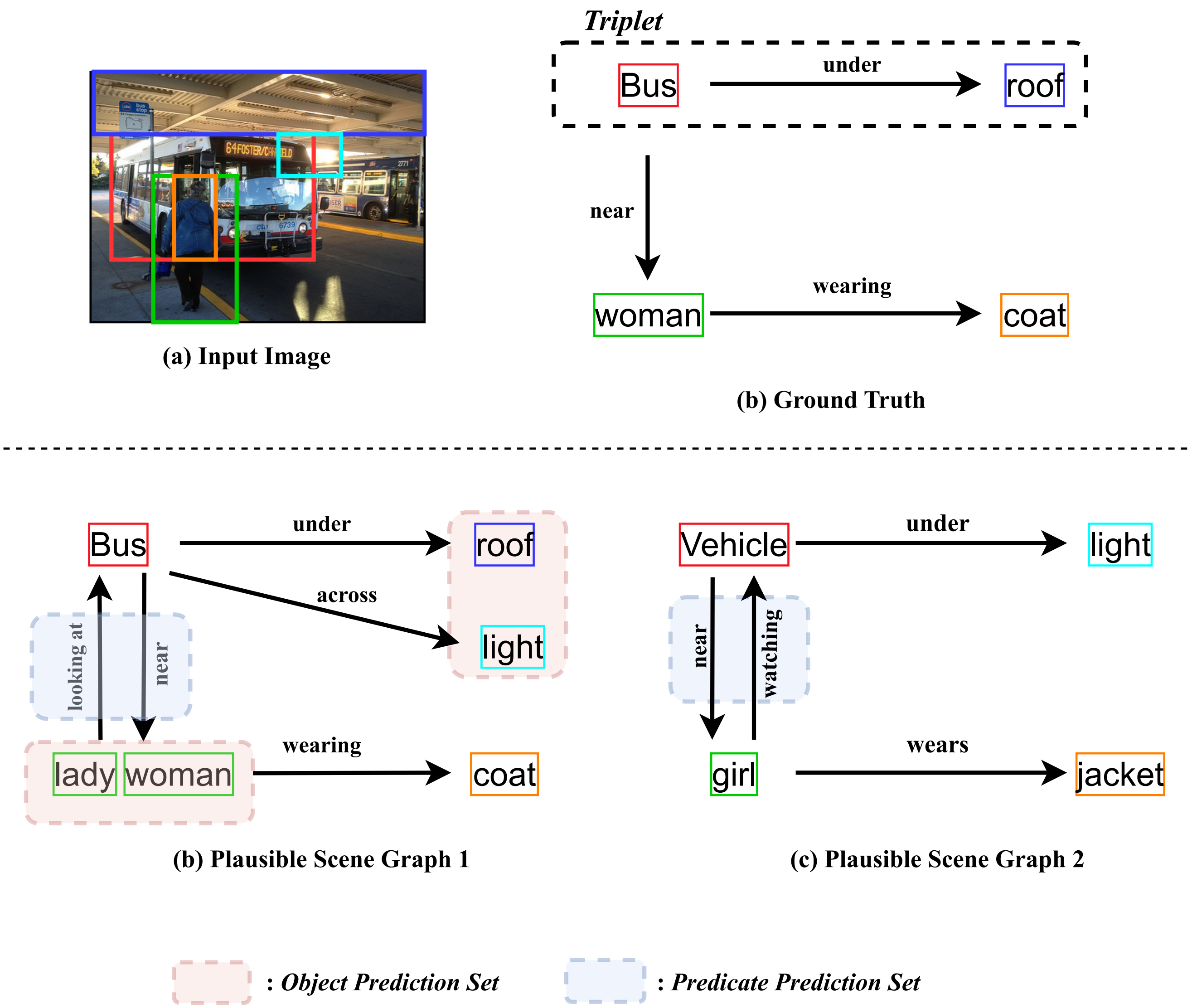
C.2.1. Impact of different prompting strategies on average set size of triplet prediction sets

In Fig 6 of the main paper, we showed how incontext learning style prompting strategy benefits the MLLM-based plausibility assessment task, of truncating triplet prediction sets without impacting their original empirical coverage. Fig II shows how changing our proposed prompting strategy impacts the average set size of the truncated prediction sets. We can observe the figure, that in general when the example prompt is not provided as a one-shot support example [2], the $AvgSize$ increases, thus highlighting its importance for our task. On the other hand, the impact of not using image cropping or the system prompt is relatively small on the empirical set size of the truncated sets. However, given their impact on the Cov_T values (Fig 6), it can be concluded that utilizing our full prompting strategy provides the most optimal results.

C.2.2. Impact of number of options in the MCQA prompt

We choose at max 5 entries from a triplet prediction set, which along with the ‘no valid option’ choice make a total of 6 options in the MCQA prompt to the MLLM. We observe empirically that 5 options give the optimal performance. This can be validated from Table I, which shows that increasing the number of options adversely affects the coverage while decreasing the number adversely affects the average set size.

This phenomenon occurs because increasing the number of options confuses the MLLM, causing it to hallucinate tokens that indicate all choices are equally plausible. This results in near-uniform likelihood values across all tokens (where each token represents an option). Consequently, the MLLM behaves like a naive or random guesser, selecting entries that are implausible and often not the ground truth. This significantly impacts empirical coverage. Conversely, limiting the number of options preserves the original empirical coverage but restricts the truncation of the prediction set. This happens because, in the task of plausibility assessment, the MLLM inherently compares the provided options. When the token space is reduced, the comparison is constrained, leading to the inclusion of more entries in the set. Thus, the optimal number of options for the MCQA prompt is 5.



(a) Input Image

(b) Ground Truth

(b) Plausible Scene Graph 1

(c) Plausible Scene Graph 2

Figure III. Qualitative results on a test image with BGNN+PC-SGG. The prediction sets obtained via PC-SGG facilitate the generation of multiple plausible scene graphs.

C.3. Qualitative Visualization

Fig III shows some plausible scene graphs generated by the BGNN+PC-SGG method. The importance of each plausible scene graph will depend on the downstream application and must be determined by domain expertise.

References

[1] Anastasios Nikolas Angelopoulos, Stephen Bates, Michael Jordan, and Jitendra Malik. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*, 2021. 3

[2] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 5

[3] Deunsol Jung, Sanghyun Kim, Won Hwa Kim, and Minsu Cho. Devil's on the edges: Selective quad attention for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18664–18674, 2023. 2

[4] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 4, 5

[5] 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. 2

[6] Shuo Li, Sangdon Park, Xiayan Ji, Insup Lee, and Osbert Bastani. Towards pac multi-object detection and tracking. *arXiv preprint arXiv:2204.07482*, 2022. 3

[7] Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591, 2020. 3

[8] Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 2019. 2, 3

[9] 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

[10] Alexander Timans, Christoph-Nikolas Straehle, Kaspar Sakmann, and Eric Nalisnick. Adaptive bounding box uncertainties via two-step conformal prediction, 2024. 1, 3

[11] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5831–5840, 2018. 2