Conformal Prediction and MLLM aided Uncertainty Quantification in Scene Graph Generation

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

Abstract

Scene Graph Generation (SGG) aims to represent visual scenes by identifying objects and their pairwise relationships, providing a structured understanding of image content. However, inherent challenges like long-tailed class distributions and prediction variability necessitate uncertainty quantification in SGG for its practical viability. In this paper, we introduce a novel Conformal Prediction based framework, adaptive to any existing SGG method, for quantifying their predictive uncertainty by constructing well-calibrated prediction sets over their generated scene graphs. These scene graph prediction sets are designed to achieve statistically rigorous coverage guarantees under exchangeability assumptions. Additionally, to ensure the prediction sets contain the most practically interpretable scene graphs, we propose an effective MLLM-based post-processing strategy for selecting the most visually and semantically plausible scene graphs within each set. We show that our proposed approach can produce diverse possible scene graphs from an image, assess the reliability of SGG methods, and improve overall SGG performance.

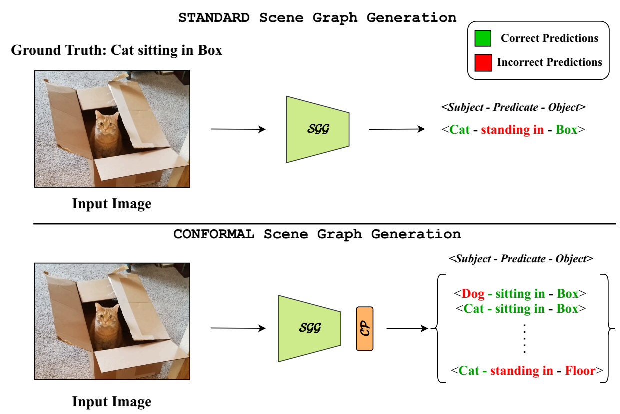


Figure 1. **Distinction between standard and conformal SGG**. The upper half shows how a standard SGG method generates a single prediction of a triplet in the image's scene graph. The lower half shows how, by adding conform prediction blocks on top of an SGG model, we can generate prediction sets for each triplet in the scene graph, which quantifies the underlying model's uncertainty and improve the chances of covering the actual ground truth.

1. Introduction

Scene graphs provide a structured semantic representation of a scene depicted in an image. The nodes and edges of this semantic graph structure represent the objects and their bipartite relationships. Such semantic representation of images bridges the gap between vision and language, playing a crucial role in numerous multimodal reasoning tasks such as visual question answering [48], image retreival [18], as well as dynamic safety-critical tasks such as autonomous mobility [31] and robot path planning [1, 35]. However, despite numerous strides made in end-to-end scene graph generation (SGG), this task remains highly challenging due to inherent ambiguities arising from multiple sources, such as imprecise scene descriptions, imperfect object detection, linguistic variations (synonyms and homonyms) [9, 13, 29, 32], long-tailed class distributions, and multiple viewpoint interpretations. As such, state-of-the-art SGG models tend to produce noisy scene graphs, raising the need to quantify their predictive uncertainty for reliable application to downstream tasks [23].

While there have been works to incorporate uncertainty quantification (UQ) in SGG for performance improvement [54], these approaches are designed specifically for their proposed model architectures. In contrast, this paper focuses on a *post-hoc* distribution-free and model-agnostic UQ method with formal statistical coverage guarantees. Specifically, we employ *Conformal Prediction* (CP) [3, 33, 50, 53] to design a model agnostic UQ framework for state-of-the-art SGG models and provide prediction sets of an image's scene graph. The prediction sets constructed via CP aim to *provably contain the true ground-truth class with a desired probability* [4]. This enables mod-

*Work done while at UCR.

eling of the inherent predictive uncertainty of an underlying pre-trained SGG model, with safety assurances [50]. We refer to this form of SGG as *Conformal SGG*, and its distinction with standard SGG is shown in Fig 1.

Since end-to-end SGG involves detecting all objects in an image and classifying their pairwise relationships, incorporating CP entails the construction of prediction sets for the object and relationship classification. Formally, this pair of objects and their relationship is called a $< subject - predicate - object >$ triplet [22], forming the fundamental block of a scene graph. Therefore, by using CP to collectively construct prediction sets for objects and predicates, we can subsequently derive a prediction set for the whole triplet. This facilitates the generation of multiple possible scene graphs for an image in the form of prediction sets with formal coverage guarantees

The main objectives of this work are to: 1) design a CP-based post-hoc model-agnostic UQ method adaptive to any existing SGG framework, enabling the generation of multiple possible scene graphs from an image that can provide diverse possible scenarios in a downstream safety-critical task such as robot path planning [34, 35], 2) assess the reliability of existing state-of-the-art SGG methods by analyzing their predictive uncertainty, and 3) show that conformal SGG improves the over recall hit rate compared to standard SGG. As previously mentioned, long-tailed class distribution is a common problem in SGG [13], as such prediction sets from standard CP, which aim to achieve aggregate coverage guarantees across the entire dataset, may fail to provide consistent coverage for each class [14, 38], particularly those in the tail of the distribution. To address this issue, we design our SGG-specific CP framework as *class-conditional* CP [38] to ensure statistical coverage guarantees at the individual class level as opposed to the dataset level.

While CP-based prediction sets of each triplet are constructed to possibly achieve desired statistical coverage, depending on the classes captured in the individual object and predicate sets, many entries in the overall triplet prediction set may be visually uncorrelated to the scene and/or semantically implausible. This affects the interpretability of the prediction sets and has the practical drawback of large set sizes. To address this issue, we propose a post-processing filtering operation based on scene coherence and feasibility conditions using Multimodal Large Language Models (MLLM) [24]. Having been trained on extensive natural image and language pairs, these models develop an implicit understanding of real-world inter-object relationships, providing a foundational prior for selecting the most *plausible* scene descriptions associated with an image. Therefore, by converting the entries in each triplet prediction set into language descriptions, we design an effective prompting strategy for in-context learning of an MLLM to achieve the desired task of selecting the most semantically plausible entries in the prediction set without compromising desired coverage guarantees. This operation improves the practicality of the prediction sets by compressing them into the most interpretable detections. We refer to the entire pipeline as **P**lausibility ensured **C**onformal **SGG**, or **PC-SGG**.

To the best of our knowledge, *this is the first work to propose a CP framework for statistically rigorous UQ for SGG, along with an MLLM-based post-processing strategy for set size minimization via plausibility assessment.* The main contributions are as follows:

- We propose a statistically rigorous post-hoc conformal prediction framework for UQ of SGG models by generating scene graph prediction sets, which facilitates the generation of diverse scene graphs from an image.
- To ensure the practical viability of the prediction sets, we propose an MLLM-based post-processing strategy for filtering prediction set entries that are visually and/or semantically implausible.
- Using PC-SGG, we assess the reliability of numerous SGG frameworks by analyzing the empirical coverage of their prediction sets.
- We further demonstrate that incorporating our approach improves the recall-hit rate compared to standard SGG.

2. Related Works

Scene Graph Generation. Scene Graph Generation (SGG) predominantly uses detection-based methods, following a two-stage framework: object detection followed by relationship detection. Visual Genome [22] provided foundational insights into the inherent ambiguities in human annotations of visual relationships. Key approaches include message passing structures [27], linguistic information integration [56], region captioning-based methods [28], low-dimensional vector mapping [59], transformer-based networks [25, 46], and feature fusion techniques [12]. Recent methods for relationship detection include iterative approaches [46, 47] that progressively refine relationship detector parameters, while others focus on predicate representation strategies, probabilistic modeling [54], fine-grained predicate learning [30], and hybrid-attention mechanisms [15]. Another direction explores sampling strategies, such as bipartite graph networks with adaptive message passing and bi-level data resampling [26]. However, the inherent challenges in SGG [13, 32] necessitate the need for a plug-in tool to assess the predictive uncertainty and, consequently, reliability of these approaches.

Uncertainty in Scene Graph Generation. Research on uncertainty in scene graph generation (SGG) has largely addressed relationship ambiguity by exploring statistical regularities in scene graphs [58], demonstrating how context influences predicate prediction. Works like [55] introduced an attentional graph convolutional network for implicit uncertainty handling in relationship proposals. Un-

biased scene graph generation is also explored in [45], addressing the uncertainty arising from long-tailed distributions in relationship detection. Methods like [17] included uncertainty modeling for highlighting semantic ambiguities in visual question answering through scene graphs. However, most of these methods provide strategies for UQ specific to their model architecture and lack safety assurances regarding ground-truth coverage. In this work, we design a model-agnostic UQ method with such statistical guarantees.
Conformal Prediction. Conformal Prediction (CP) provides a framework for producing prediction sets with guaranteed coverage under the assumption of exchangeability [52]. The method has gained significant traction in machine learning applications due to its distribution-free nature and rigorous statistical guarantees. In computer vision, CP has been successfully applied to various tasks: [4] utilized CP for image classification uncertainty quantification, while [2] extended it to object detection tasks. The framework has also shown promise in causal inference [49] and time series forecasting [43]. In the context of deep learning, CP has been integrated with neural networks to provide uncertainty estimates in high-stakes applications [21]. However, the application of CP to structured prediction tasks like scene graph generation remains largely unexplored, particularly when dealing with the challenge of large prediction sets while maintaining coverage guarantees.

3. Preliminaries

3.1. Conformal Prediction

Assume a dataset $\mathcal{D} = \{\mathcal{D}_{tr}, \mathcal{D}_{cal}, \mathcal{D}_{test}\}$, comprised of a training set $\mathcal{D}_{tr}$, a calibration set $\mathcal{D}_{cal} = \{(X_i, Y_i)\}_{i=1}^{n} \sim P_{XY}$, and a test set $\mathcal{D}_{test} = \{(X_j, Y_j)\}_{j=n+1}^{n+n_t} \sim P_{XY}$. Now given a pre-trained model, f, for an unseen test sample X_{n+1}, generalized CP [6, 33] entails the construction of a prediction set $\hat{\mathcal{C}}(X_{n+1}) = \text{SET}(f(X_{n+1}), \mathcal{A}, \mathcal{D}_{cal})$, where $\text{SET}(\cdot)$ is a function that constructs sets from the outputs $f(X_{n+1})$ based on $\mathcal{D}_{cal}$ and a *nonconformity measure* $\mathcal{A}$. Under the assumption of exchangeability (relaxed i.i.d) [50], i.e. $\mathcal{D}_{cal} \cup (X_{n+1}, Y_{n+1})$, $\hat{\mathcal{C}}(X_{n+1})$ attains the following *marginal* probabilistic guarantee,

$$P(Y_{n+1} \in \hat{\mathcal{C}}(X_{n+1})) \geq 1 - \alpha \tag{1}$$

for some tolerated *miscoverage* rate α [41]. This marginal guarantee provided by generalized CP holds in average across $\mathcal{D}_{test}$ provided $\mathcal{D}_{cal}$ and $\mathcal{D}_{test}$ came from a fixed distribution [50]. In contrast, if granular guarantees are required (for example, at class level or cluster level), *conditionally* valid coverage is necessary [19, 37, 40].In this work, we are particularly interested in *class-conditional* coverage [8, 39, 53], yielding the following statistical guarantee, for a test sample X_{n+1},

$$P(Y_{n+1} \in \hat{\mathcal{C}}(X_{n+1}) \mid Y_{n+1} = y) \geq 1 - \alpha_y \quad \forall y \in \mathcal{Y} \tag{2}$$

where, $\mathcal{Y} = \{1, \ldots, K\}$ are distinct class-labels for a multi-class classification problem, and $\{\alpha_y\}_{y \in \mathcal{Y}}$ are classwise miscoverage rates. Eq. 2 implies Eq. 1 and provides a stronger guarantee as it controls coverage levels within each class [14, 38, 50].

3.2. Scene Graph Generation

SGG entails obtaining a graph-structured representation, $G_t = \{S, R, O\}$, of an image I, describing the atomic inter-object interactions of the scene in I. It is achieved by detecting and subsequently classifying all the objects in a scene and then classifying their pairwise relationships. Here $S = \{s_1, s_2, ..., s_{N_o}\}$ and $O = \{o_1, o_2, ..., o_{N_o}\}$ are respectively the subject and object list both of which map to the same list of N_o objects present in I, and $R = \{r_1, r_2, ..., r_{N_r}\}$ is the list of bipartite object relationships called predicates. Thus, G_t is obtained by combinatorially arranging subject-object pairs into triplets $(< subject - predicate - object >)$, $t_{ij} = (s_i, r_k, o_j)$. We refer to the set of object and predicate classes as $\mathcal{Y}_o = \{y_1^o, y_2^o, ..., y_{K_o}^o\}$ and $\mathcal{Y}_r = \{y_1^r, y_2^r, ..., y_{K_r}^r\}$ respectively.

4. Method

The overview of PC-SGG is shown in Figure 2. Given a pre-trained SGG model ϕ, we first apply class-conditional CP to the object and predicate classifier heads (f_o, f_r) and obtain prediction sets for the objects and predicates, respectively. The two prediction sets are then combined to obtain a prediction set for each detected triplet. Each of these triplet prediction sets is then converted into sets of language descriptions to be assessed by an MLLM-based post-processing unit. Leveraging a question-answer strategy, the MLLM assesses each of these descriptions and compresses each triplet prediction set into the set of most plausible triplets. In the subsequent sections, we describe in detail the conformal prediction setup for SGG and the MLLM-based post-processing method.

4.1. Conformal Scene Graph Generation

For any test sample X_{n+1} we denote the object and predicate prediction sets as $\hat{\mathcal{C}}_o(X_{n+1})$ and $\hat{\mathcal{C}}_r(X_{n+1})$ respectively. To prevent confusion, it must be noted that a sample, X, for the case of the objects, refers to the region of interest (RoI) [36], and for the case of predicates (and triplets), refers to a pair of detected objects which, among other information, includes the RoI associated with union box of the pair [20, 26, 57]. Inspired by [39], we design the overall CP framework as a class-conditional prediction set classifier. The overall algorithm is shown in Algorithm 1 (refer to supplementary A.1).
Conformal Calibration. In order to achieve the nominal class-conditional coverage of Eq. 2, well-calibrated prediction sets need to be constructed to measure how well an

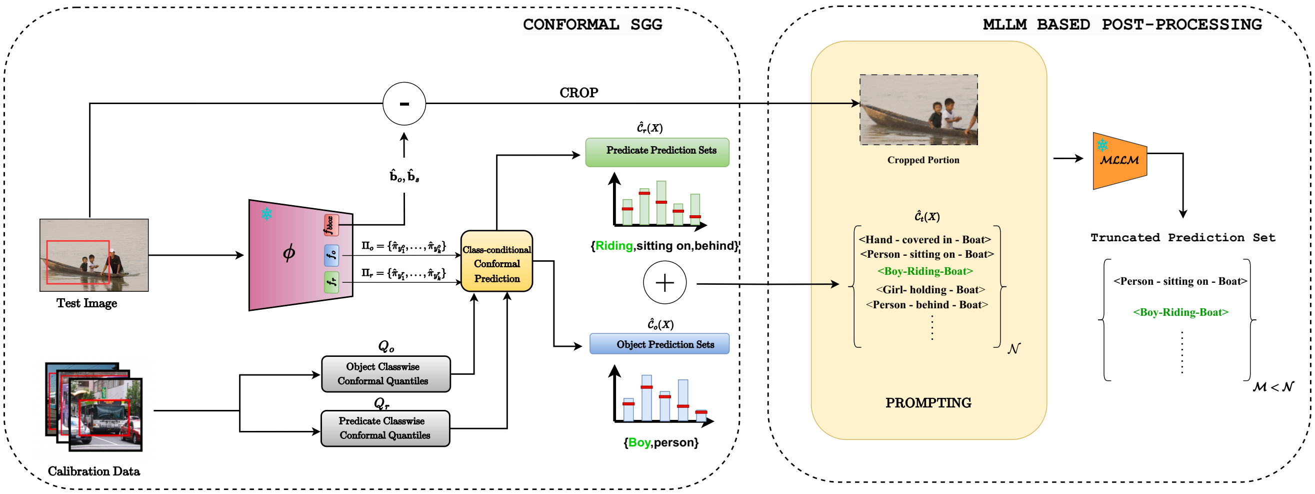


Figure 2. **Overview of PC-SGG Pipeline.** For each *Test Image*, a pre-trained SGG model, ϕ, is used to obtain object bounding boxes $\hat{\mathbf{b}}_o$ (using f_{bbox}), object classification probabilities Π_o (using f_o), and the probabilities of their pairwise predicates classifications which include the classification Π_r (using f_r). Using quantiles (Q_o, Q_r) derived from a *Calibration Data*, we construct class-conditional conformal sets for both objects ($\hat{\mathcal{C}}_o(X)$) and predicates ($\hat{\mathcal{C}}_r(X)$). These conformal sets are then combinatorially combined ($\oplus$) to generate a *Triplet Prediction Set*, $\hat{\mathcal{C}}_t(X)$. To assess the plausibility of each entry of $\hat{\mathcal{C}}_t(X)$, we leverage an MLLM-based post-processing unit. The entries of $\hat{\mathcal{C}}_t(X)$ are converted into textual descriptions, which, along with the cropped portion of the test image defined for the triplet set (cropped using the union bounding box of the triplet's object pairs) is converted into an input prompt for the MLLM to process and predict the *truncated prediction set* of the most plausible triplets as a next token prediction problem.

unseen test sample *conforms* to the training distribution. As such, using an appropriate non-conformity measure, $\mathcal{A}$, classwise non-conformity scores are obtained on $\mathcal{D}_{cal}$, which are then used to obtain class-wise conformal quantile values, $\hat{q}_y$, that act as the necessary thresholds for assessing the conformity of any test sample. In our setup, we define the non-conformity measure as $\mathcal{A}(f(X), y) = 1 - \hat{\pi}_y$, where $\hat{\pi}_y$ is the softmax probability output for class y such that $f(X) = \{\hat{\pi}_y\}_{y\in\mathcal{Y}}$.

As shown in Algorithm 1, calibration proceeds by first greedily matching each pair of predicted objects to ground truth as follows,

$$match(\hat{\mathbf{b}}_s, \mathbf{b}_s^j, \hat{\mathbf{b}}_o, \mathbf{b}_o^j) = \underset{j}{\text{argmax}}\left(\frac{1}{2}(\Gamma(\hat{\mathbf{b}}_s, \mathbf{b}_s^j) + \Gamma(\hat{\mathbf{b}}_o, \mathbf{b}_o^j))\right)$$
$$\text{s.t.} \quad \Gamma(\hat{\mathbf{b}}_s, \mathbf{b}_s^j) \geq 0.5 \;\wedge\; \Gamma(\hat{\mathbf{b}}_o, \mathbf{b}_o^j) \geq 0.5 \tag{3}$$

where Γ is the Intersection over Union (IoU) function, $\hat{\mathbf{b}}_s, \hat{\mathbf{b}}_o \in \mathbb{R}^4$ are the predicted subject-object pair's bounding boxes, $\mathbf{b}_s^j, \mathbf{b}_o^j \in \mathbb{R}^4$ are the bounding boxes of the j^{th} ground-truth subject-object pair. Using the matched ground truth classes (objects and predicate), the class-specific non-conformity score is computed for both the objects and the predicate. Therefore, for the k^{th} object class, the aggregated list of non-conformity scores is obtained as $S_{y_k^o} = \{\mathcal{A}(f_o(X_i), Y_k^o)\}_{i=1}^{n_{y_k^o}} = \{s_i^o\}_{i=1}^{n_{y_k^o}}$, where $y_k^o \in \mathcal{Y}_o$. Similarly for the k^{th} predicate class we get $S_{y_k^r} = \{\mathcal{A}(f_o(X_i), Y_k^r)\}_{i=1}^{n_{y_k^r}} = \{s_i^r\}_{i=1}^{n_{y_k^r}}$ where $y_k^r \in \mathcal{Y}_r$. Finally, the class-specific conformal quantiles are obtained as follows,

$$\begin{aligned} \hat{q}_{y_k^o} &= \left\lceil (n_{y_k^o}+1)(1-\alpha_{y_k^o})/n_{y_k^o} \right\rceil \text{-th empirical quantile of } S_{y_k^o}. \\ \hat{q}_{y_k^r} &= \left\lceil (n_{y_k^r}+1)(1-\alpha_{y_k^r})/n_{y_k^r} \right\rceil \text{-th empirical quantile of } S_{y_k^r}. \end{aligned} \tag{4}$$

The list of class-wise conformal quantiles for the objects and predicates is denoted as $Q_o = \{\hat{q}_{y_k^o}\}_{y_k^o\in\mathcal{Y}_o}$ and $Q_r = \{\hat{q}_{y_k^r}\}_{y_k^r\in\mathcal{Y}_r}$, respectively. For simplicity, we set $\alpha_{y_k^o} = \alpha_o \;\forall\; y_k^o \in \mathcal{Y}_o$ and $\alpha_{y_k^r} = \alpha_r \;\forall\; y_k^r \in \mathcal{Y}_r$.
Conformal Inference. During inference, utilizing Q_o and Q_r, $\hat{\mathcal{C}}_o(X_{n+1})$ and $\hat{\mathcal{C}}_r(X_{n+1})$ are obtained as follows,

$$\begin{aligned} \hat{\mathcal{C}}_o(X_{n+1}) &= \{y_k^o \in \mathcal{Y}_o : \hat{\pi}_{y_k^o} \geq 1 - \hat{q}_{y_k^o}\} \\ \hat{\mathcal{C}}_r(X_{n+1}) &= \{y_k^r \in \mathcal{Y}_r : \hat{\pi}_{y_k^r} \geq 1 - \hat{q}_{y_k^r}\} \end{aligned} \tag{5}$$

where the meaning of an unseen test sample X_{n+1}, for the object and predicate cases is defined earlier. As shown in Fig 2, $\hat{\mathcal{C}}_o(X_{n+1})$ and $\hat{\mathcal{C}}_r(X_{n+1})$ are combinatorially aggregated to obtain the prediction set for the entire triplet $\hat{\mathcal{C}}_t(X_{n+1})$. The strength of the coverage guarantees for the object and predicate prediction sets control the overall coverage guarantee for the triplet.

Theorem 1. *Given the ground truth class of the k^{th} triplet is denoted as $y_k^t = [y_k^s, y_k^r, y_k^o] \in \mathbb{R}^3$ where $y_k^s, y_k^o \in \mathcal{Y}_o$ and $y_k^r \in \mathcal{Y}_r$, the triplet coverage guarantee is given as $P(y_k^t \in \hat{\mathcal{C}}_t(X_{n+1}^r)) = P(Y_{n+1}^o \in \hat{\mathcal{C}}_o(X_{n+1}^o) \mid Y_{n+1}^o = y_i^o) \cdot P(Y_{n+1}^r \in \hat{\mathcal{C}}_o(X_{n+1}^r) \mid Y_{n+1}^r = y_m^r) \;\forall\; y_k^s \in \mathcal{Y}_o, y_k^o \in \mathcal{Y}_o, y_k^r \in \mathcal{Y}_r$.*

Please see Proof in supplementary.

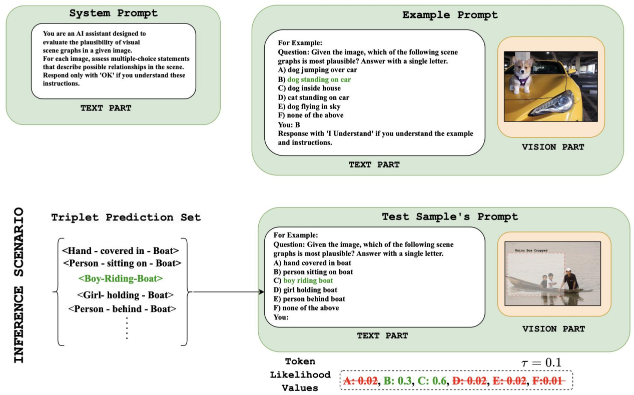


Figure 3. **Prompting strategy for plausibility assessment.** First, a *system prompt* outlines the task for the MLLM. Then, an *example prompt* is created with a randomly sampled image from the calibration set and a hand-crafted text description, framing plausibility assessment as an MCQA problem. During inference, entries from a test image's triplet prediction set are processed in groups of 5, for designing the MCQ text prompts similar to the example. The vision part of the prompt is the cropped portion of the test image linked to the detected triplet. The MLLM's token likelihoods are thresholded by τ to identify the most plausible choices for the scene. The ground-truth triplet is highlighted in green for both the example and inference scenarios.

Corollary 1. *Following Theorem 1,* $P(y_k^t \in \hat{\mathcal{C}}_t(X_{n+1}^r)) \geq (1-\alpha_o)(1-\alpha_r)$, $\forall\, y_k^s \in \mathcal{Y}_o, y_k^o \in \mathcal{Y}_o, y_k^r \in \mathcal{Y}_r$.

To ensure a nominal coverage of each triplet prediction set to be approximately close to 90%, we choose $\alpha_o = 0.05$ and $\alpha_r = 0.1$ in our experiments.

4.2. MLLM Guided Plausibility Assessment

While the triplet prediction sets constructed using conformal SGG have nominal coverage guarantees, in practice, many entries in a triplet prediction set may not be plausible. For example, for the image shown under the *Inference Scenario* in Fig 3, its triplet prediction set has entries like "`hand covered in boat`, `girl holding boat`". The first description of the triplet is implausible from a practical linguistic/semantic perspective, while the second one does not correlate to the actual scene in the image and is visually implausible. This causes the overall prediction set of the triplet to be bloated with impractical and uninterpretable scene descriptions. We propose a plausibility checking unit which addresses this issue by leveraging the foundational prior of an MLLM in associating the right scene description to the scene at hand. The overview of this unit's functioning is shown in Fig 3.

Plausibility as multiple-choice Q&A. We craft an effective prompting strategy that query's an MLLM regarding the plausibility of an entry in a triplet prediction set. For any detected triplet in a test image, we have the predicted bounding boxes for its subject, $\hat{\mathbf{b}}_s$, and object, $\hat{\mathbf{b}}_o$, as well as, the prediction set $\hat{\mathcal{C}}_t(X)$ obtained via the conformal prediction. First, we convert the entries of $\hat{\mathcal{C}}_t(X)$ into language descriptions, then sequentially sample 5 language descriptions to create a multiple choice question (MCQ), forming the *text part* of the prompt to the MLLM (Fig 3). The *vision part* of the prompt is composed of the cropped-out portion of the test image defined by the union box $\hat{\mathbf{b}}_s \bigcup \hat{\mathbf{b}}_o$. The MLLM is then tasked with answering the MCQ, thereby choosing the most plausible scene description among the 5 choices. This boils down the next-token prediction of the MLLM into a single-token prediction, with the token space being restricted to the number of choices in the MCQ [16, 42]. In practice, we add a 'no valid option' choice to MCQ as shown in Fig 3. In this way, we task the MLLM to eliminate the implausible entries in each $\hat{\mathcal{C}}_t(X_{n+1})$ and compress it into a set of practically interpretable scene descriptions.

In order to ensure the MLLM adheres to the task at hand, we leverage a one-shot in-context learning setup [7] and provide adequate task-specific information to the MLLM in the form of an instructional *system prompt* and a toy scenario of the task at hand in the form of an *example prompt* (Fig 3). The *example prompt* contains a randomly sampled calibration set image along with a hand-crafted toy MCQA, containing the ground-truth triplet as one of the options.

Extraction of plausible scene descriptions. The *system prompt* and the *example prompt* are passed as input to the MLLM just once, while the *Inference Scenario* is run iteratively over all triplet prediction sets across all test images. The prompt of each iterative sample is essentially the cropped portion of a test image associated with the detected triplet and 5 descriptions from that triplet's prediction set used to make the MCQ. Since the token space is restricted to the number of choices in the MCQ plus the 'no valid option' choice, we obtain likelihood values for a total of 6 tokens from the MLLM. We then threshold these likelihood values using a token threshold, τ, (Fig 3), to choose the most plausible options and aggregate them as the final truncated triplet prediction set. If the 'no valid option' has the highest likelihood value, the group of 5 prediction set entries is wholly disregarded. The choice of τ is very important for this operation. This is because the token likelihood values from the MLLM may be high for certain choices, which may be plausible w.r.t. the scene but may not be the actual ground truth triplet. So, choosing a high value for τ may result in the omission of the ground truth (provided it is covered by the original prediction set) from the truncated prediction set, consequently affecting nominal coverage.

5. Experiments

5.1. Dataset and Splits

We perform experiments on the benchmark image scene graph generation dataset Visual Genome (VG) [22], composed of 108k images. To filter out the predicate classes with an extremely low number of samples we adopt the commonly used VG split called VG150 [28, 57], composed of the most frequent 150 object and 50 predicate classes.

Training and Calibration Sets. The calibration set is carved out from the 75k images in the original VG150 training and validation sets. The calibration set is designed such that its images cover at most 10% of the total number of samples in each of the object and predicate classes, ensuring that a minimum of 2 samples from any class is included in the calibration set. The final calibration set contains 7174 images. The remaining images are used for training.

5.2. Implementation and Evaluation Metrics

We incorporate PC-SGG with existing SOTA SGG methods and quantify their uncertainty for end-to-end SGG, formally called SGDET (joint detection and classification of objects and pairwise predicates). For all object classes, we set, α_o to 0.05, and for all predicate classes we set α_r to 0.1, This maintains the combined marginal coverage probability, $(1-\alpha_o)(1-\alpha_r)$, is close to 90%. The value of τ is set to 0.1. All experiments were performed using 4 NVIDIA RTX-3090.

SGG Models. We implement PC-SGG over the following 5 SOTA SGG frameworks, MOTIFS [57], MOTIFw/DEBIAISNG (MOTIFS-D) [11], VCTREE [44], BGNN [26], and SQUAT [20]. For all models, Faster R-CNN [36] with ResNeXt-101-FPN backbone is used as the object detector. The hyperparameters specific to each model are kept the same as reported in their respective papers. Since the VG150 training set is different from the usual one, we retrain the object detector and all the models from scratch on the modified training set for our experiments.

MLLM. We use BLIP-2 [24] with FLAN-T5-XL [10] as our MLLM. Due to computational resource constraints, we restrict our experiments to the 8-bit quantized model. Additional implementation details are in the supplementary.

Conformal Prediction Metrics. The following CP metrics [5] are used to gauge the uncertainty associated with both object and predicate classification of each SGG model,

1. *Coverage Validity*: Computes empirical coverage of the CP procedure and checks its deviation from overall nominal coverage guarantees [50, 51],

$$Cov = 100 \times \frac{1}{n_t}\sum_{j=n+1}^{n+n_t} \mathbb{1}[Y_j \in \hat{\mathcal{C}}(X_j)] \tag{6}$$

   where, $\mathbb{1}[\cdot]$ is the indicator function.

2. *Average Class Coverage Gap*: The following metric is specific to class-conditional coverage as it measures class-wise deviation from its desired coverage level [14],

Table 1. Comparison of CP metrics of different SGG methods with PC-SGG. The Cov_T values shown here are computed over the truncated triplet set obtained after MLLM-based post-processing. The best results are highlighted in bold.

Method	Objects			Predicates			Triplets
	$Cov\uparrow$	$CovGap\downarrow$	$AvgSize\downarrow$	$Cov\uparrow$	$CovGap\downarrow$	$AvgSize\downarrow$	$Cov_T\uparrow$
MOTIFS [57]	88.94	5.8	4.87	84.11	6.2	16.09	74.97
MOTIFS-D [11]	88.94	5.8	4.87	86.67	5.9	16.81	76.67
VCTREE [44]	89.38	5.7	**4.23**	88.61	5.9	16.41	80.06
SQUAT [20]	90.26	4.9	4.48	**90.25**	**4.6**	**14.48**	80.25
BGNN [26]	**90.35**	**4.8**	4.48	89.68	5.2	16.23	**80.45**

$$CovGap = 100 \times \frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}} ||\hat{c}_y - (1-\alpha_y)||_1 \tag{7}$$
$$\hat{c}_y = \frac{1}{|\mathcal{D}^y|}\sum_{j\in\mathcal{D}^y} \mathbb{1}[Y_i \in \hat{\mathcal{C}}(X_i)]$$

   where, $\mathcal{D}^y \subset \mathcal{D}^{test}$ are all samples belonging to class y, $||\cdot||_1$ is the ℓ_1 distance, α_y is the class specific marginal error rate. For all object classes $\alpha_y = \alpha_o$ and for all predicate classes $\alpha_y = \alpha_r$. Individually for the object and predicate classes

3. *Average Set Size*: The following computes sharpness of the computed prediction sets,

$$AvgSize = \frac{1}{n_t}\sum_{j=n+1}^{n+n_t} |\mathcal{C}(X_j)| \tag{8}$$

4. *Triplet Coverage Validity*. The Cov metric is used to quantify CP performance for the object and predicate classifications individually. To assess the validity of the prediction sets associated of the whole triplet, we introduce the following metric to gauge the empirical coverage of the whole triplet prediction set,

$$\begin{aligned} Cov_T &= 100 \times \frac{1}{n_t}\sum_{j=n+1}^{n+n_t} m_j^s \wedge m_j^r \wedge m_j^o \\ m_j^s &= \mathbb{1}[Y_j^s \in \hat{\mathcal{C}}_o(X_j^s)], \quad m_j^r = \mathbb{1}[Y_j^r \in \hat{\mathcal{C}}_r(X_j^r)], \\ m_j^o &= \mathbb{1}[Y_j^o \in \hat{\mathcal{C}}_o(X_j^o)] \end{aligned} \tag{9}$$

   where, $\hat{\mathcal{C}}_o(X_j^s)$, $\hat{\mathcal{C}}_r(X_j^r)$, and $\hat{\mathcal{C}}_o(X_j^o)$ are the prediction sets of the subject, predicate, and object in the triplet, $Y_j^s, Y_j^o \in \mathcal{Y}_o$, and $Y_j^r \in \mathcal{Y}_r$, and n_t here runs over all the ground-truth triplet samples.

SGG Metrics. The common metrics used for quantifying SGG performance are Recall@K (R@K) and mean-Recall@K(mR@K). However, since PC-SGG provides prediction sets for each detected triplet, we modify the computation of the recall hit rate. Unlike R@K where the equality of the subject, predicate, and object classes in a detected triplet is checked to that of a matched ground truth, we check if the matched ground truth *is present in*

Table 2. Comparison of empirical coverage and average set size of triplet prediction sets with and without MLLM-based post-processing.

Method	w/o MLLM Plausibility Assessment		w/ MLLM Plausibility Assessment	
	$Cov_T \uparrow$	$AvgSize \downarrow$	$Cov_T \uparrow$	$AvgSize \downarrow$
MOTIFS [57]	74.97	866.09	74.97	403.21
MOTIFS-D [11]	76.93	893.21	76.67	411.58
VCTREE [44]	80.06	818.76	80.06	389.24
SQUAT [20]	80.43	816.68	80.25	398.67
BGNN [26]	80.45	971.69	80.45	464.11

the triplet prediction set obtained from PC-SGG. We refer to these metrics, as coverage-Recall@K (cR@K) and coverage-mean-Recall@K (cmR@K) which are specific to Conformal SGG and can be considered equivalent to standard SGG metrics R@K and mR@K.

5.3. Comparative Uncertainty Quantification of SGG Methods

The empirical coverage of the object, predicate, and triplet prediction sets, obtained by PC-SGG for each SGG method is shown in Table 1. For all the methods, the object prediction sets fail to achieve the class-conditional coverage guarantee of 95% (Eq 2). In comparison, the predicate prediction sets of VCTREE, SQUAT, and BGNN come close to the coverage guarantee of 90%. However, for all the methods, $AvgSize$ of the predicate prediction sets is significantly higher than that of the objects. This shows that predicate classification uncertainty is significantly higher than object classification. As such, the predicate prediction sets are larger to meet coverage guarantees. Overall, SQUAT and BGNN achieve the best empirical coverage for their triplet prediction sets, with SQUAT having the smallest $AvgSize$ for its predicate prediction sets. From the $CovGap$ values, it can also be observed that the predicate prediction sets of SQUAT achieve the lowest class-wise deviation from the class-conditional coverage guarantee. For the object prediction sets, BGNN has lowest $CovGap$. The empirical coverage values of the triplet prediction set, Cov_T, also verify Theorem 1 (check supplementary) since it is approximately close to the product of the object and predicate prediction sets' Cov values. However, the Cov_T values of none of the models achieve the marginal coverage guarantee of $(1-\alpha_o)(1-\alpha_r)$ shown in Corollary 1. This observation highlights the high predictive uncertainty of existing SGG methods, as well as, shed light on possible distribution shifts in the train and test splits of SGG datasets that make computing well-calibration conformal quantiles very challenging under exchangeability assumptions.

5.4. Results of MLLM-based Post-Processing

Table 2 shows how the average set size and empirical coverage values of the whole triplet prediction sets are impacted by applying our proposed MLLM-based plausibility assessment method as a post-processing unit. It is observed that our proposed post-processing strategy is indeed effective in significantly shrinking the size of the original triplet prediction set constructed from the conformal prediction. Additionally, in most cases, the empirical coverage Cov_T remains unaffected; only for SQUAT and MOTIFS-D a slight decrease in empirical coverage is obsevered. However, this small decrement in empirical coverage is tolerable given $> 50\%$ shrinkage in the sizes of their respective triplet prediction sets. Although the triplet prediction sets of BGNN have the highest empirical coverage, it must also be noted its prediction sets are also of the largest size since its predicate classifier is comparatively more uncertain than the other methods, as evident from Table 1.

Table 3. Comparison of cR@K and cmR@K of the different SGG methods when PC-SGG is added. Performance of these methods for standard SGG based on R@K and mR@K are also shown.

Method	**R@50**	**R@100**	**mR@50**	**mR@100**
MOTIFS [57]	23.61	29.08	4.52	6.22
MOTIFS-D [11]	24.33	30.12	5.26	7.06
VCTREE [44]	26.77	31.46	5.73	7.14
SQUAT [20]	26.81	32.06	9.95	12.05
BGNN [26]	30.07	34.90	9.63	11.92
Method+PC-SGG	**cR@50**	**cR@100**	**cmR@50**	**cmR@100**
MOTIFS [57]	38.45	46.79	25.49	34.03
MOTIFS-D [11]	40.21	47.46	26.17	35.63
VCTREE [44]	41.89	49.90	27.84	36.75
SQUAT [20]	43.23	51.87	30.94	39.23
BGNN [26]	**46.32**	**53.81**	**32.52**	**40.36**

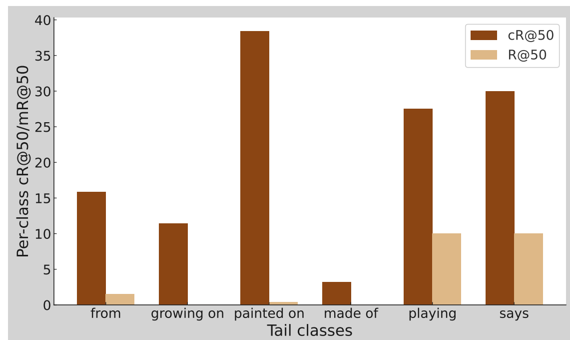
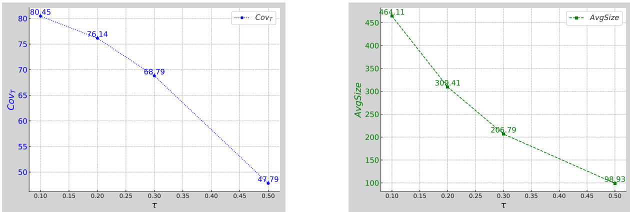


Figure 4. The R@50 and cR@50 for 5 of tail classes of VG150, with the least number of samples. The results are for the BGNN model. It can be observed that the triplet prediction sets from BGNN+PC-SGG, significantly improve the performance of detecting the tail class predicates compared to the standalone model.

5.5. Comparsion of SGG Metrics after incorporating PC-SGG

From Table 3 one can observe that BGNN+PC-SGG achieves the highest cR@K and cmR@K values, showcasing the best *conformal SGG* performance across all the SGG models used in this paper. When PC-SGG is added, the



(a) Cov_T vs τ (b) $AvgSize$ vs τ

Figure 5. Comparison of Cov_T and $AvgSize$ as functions of token threshold, τ, for BGNN+PC-SGG.

recall-hit rate improves significantly since the set-valued triplet predictions provide a much greater chance of covering the actual ground truth than the standard SGG scenario, where each detected triplet has only one prediction instead of a set of predictions. As observed in Table 3, the recall-hit rate improves by an average of 15.70 % points when comparing cR@50 with R@50.

Also, since PC-SGG is designed to achieve class-conditional coverage guarantees, adding it to any SGG method significantly improves performance on the tail classes. Across all SGG methods, we can observe that the class-conditional recall-hit rate improves by an average of 21.57 % points when comparing cmR@50 with mR@50. Note that the R@K and mR@K values of the standalone models differ from those reported in their respective papers since we retrain them on a new training split of VG150. Some qualitative results are provided in the supplementary.

5.6. Analysis

We analyze different designs for the MLLM based post-processing strategy for the BGNN model's prediction sets.

Impact of token threshold. As mentioned in Sec 4.2, the choice τ significantly affects the empirical coverage, as well as, the size of the truncated triplet prediction sets. As observed in Fig 5 varying τ between $\{0.1, 0.2, 0.3, 0.5\}$ shows when tau, although the $AvgSize$ decreases, it comes at the cost of a drastic fall in empirical triplet coverage. This is because, in multiple scenarios, the tokens predicted by the MLLM with very high probabilities may map to plausible descriptions of the scene in the test image but may not be the actual ground truth scene description. Consequently, the truncated prediction sets omit the ground truth itself. Therefore, for all experiments, we choose $\tau = 0.1$.

Impact of image cropping. From Fig 6, we can observe that if the test image is not cropped with the union feature of the subject and object bounding boxes, the empirical coverage drops. This is inherent because the visual information is not restricted to the triplet but involves other background information that distorts the MLLM's perception, resulting in the prediction set entries that have a low correlation with the cropped scene being incorrectly chosen to be part of the truncated prediction sets.

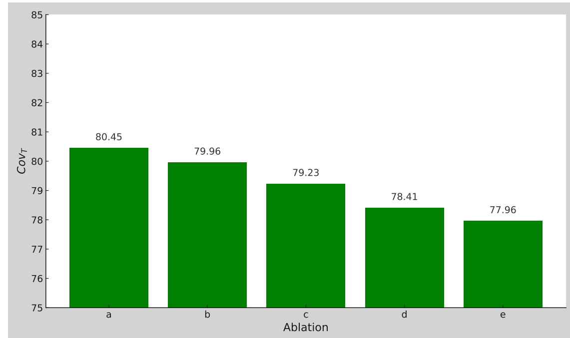


Figure 6. **Impact of different prompting strategies on Cov_T**, (a)Proposed prompting design, (b) No Image Cropping, (c) No System Prompt, (d) No Example prompt, (e) No System and example prompt. All results are obtained for the BGNN model.

Impact of in-context learning. By comparing the empirical triplet coverage values of (a) with (c), (d), and (e) shown in Fig 6, we can observe that the System and Example Prompts are critical for the MLLM's understanding of the task at hand, as without the system prompt and the one-shot in-context example there is a significant drop in empirical coverage. In fact, we observe that without the in-context example, the token prediction space does not remain restricted to the MCQ choices, and random tokens are also predicted. More analysis is provided in the supplementary.

6. Conclusion

We present a novel post-hoc model agnostic method for quantifying the uncertainty in scene graph generation methods, such that their reliability can be gauged for safety-critical applications. To this end, we propose **P**lausibility ensured **C**onformal **SGG** or **PC-SGG**. PC-SGG is comprised of a novel conformal prediction framework specific to SGG that facilitates uncertainty quantification by providing prediction sets of the possible scene graphs in an image. To ensure the plausibility of the scene graphs in each prediction set, we also propose a novel MLLM-based plausibility assessment strategy as a post-processing unit of PC-SGG. We introduce an effective MCQA style prompting strategy for tasking an MLLM to assess the plausibility of the scene graphs and truncate the prediction sets. We show how PC-SGG quantifies the predictive uncertainty of numerous SGG methods. We also show that utilizing PC-SGG significantly improves SGG performance as the generation of multiple plausible scene graphs in an image improves the chances of capturing the ground-truth graph. In future work, we would extend the method to video scene graph generation.

Acknowledgments. This work was partially supported by NSF grants OAC-2411453, CCF-2008020, CNS-2312395, CMMI-2326309 and DURIP N000141812252.

References

[1] Christopher Agia, Krishna Murthy Jatavallabhula, Mohamed Khodeir, Ondrej Miksik, Vibhav Vineet, Mustafa Mukadam, Liam Paull, and Florian Shkurti. Taskography: Evaluating robot task planning over large 3d scene graphs. In *Conference on Robot Learning*, pages 46–58. PMLR, 2022. 1

[2] Leo Andeol, Thomas Fel, Florence de Grancey, and Luca Mossina. Confident object detection via conformal prediction and conformal risk control: an application to railway signaling. In *Proceedings of the Twelfth Symposium on Conformal and Probabilistic Prediction with Applications*, pages 36–55. PMLR, 2023. 3

[3] Anastasios N Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 2023. 1

[4] 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. 1, 3

[5] Anastasios N Angelopoulos, Stephen Bates, et al. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591, 2023. 6

[6] Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 2021. 3

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

[8] Maxime Cauchois, Suyash Gupta, and John C Duchi. Knowing what You Know: Valid and validated confidence sets in multiclass and multilabel prediction. *Journal of Machine Learning Research*, 2021. 3

[9] 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. 1

[10] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024. 6

[11] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019. 6, 7

[12] Bo Dai, Yuqi Zhang, and Dahua Lin. Detecting visual relationships with deep relational networks. In *Proceedings of the IEEE conference on computer vision and Pattern recognition*, pages 3076–3086, 2017. 2

[13] Alakh Desai, Tz-Ying Wu, Subarna Tripathi, and Nuno Vasconcelos. Learning of visual relations: The devil is in the tails. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15404–15413, 2021. 1, 2

[14] Tiffany Ding, Anastasios Angelopoulos, Stephen Bates, Michael Jordan, and Ryan J Tibshirani. Class-conditional conformal prediction with many classes. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3, 6

[15] Xingning Dong, Tian Gan, Xuemeng Song, Jianlong Wu, Yuan Cheng, and Liqiang Nie. Stacked hybrid-attention and group collaborative learning for unbiased scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19427–19436, 2022. 2

[16] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2020. 5

[17] Drew A. Hudson and Christopher D. Manning. Gqa: a new dataset for compositional question answering over real-world images. *ArXiv*, abs/1902.09506, 2019. 3

[18] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3668–3678, 2015. 1

[19] Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Batch Multivalid Conformal Prediction. *International Conference on Learning Representations*, 2023. 3

[20] 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. 3, 6, 7

[21] Benjamin Kompa, Jasper Snoek, and Andrew Beam. Second opinion needed: communicating uncertainty in medical machine learning. *npj Digital Medicine*, 4, 2021. 3

[22] 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. 2, 5

[23] Dachuan Li, Bowen Liu, Zijian Huang, Qi Hao, Dezong Zhao, and Bin Tian. Safe motion planning for autonomous vehicles by quantifying uncertainties of deep learning-enabled environment perception. *IEEE Transactions on Intelligent Vehicles*, 2023. 1

[24] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2, 6

[25] Ping Li, Zhou Yu, and Yibing Zhan. Deep relational self-attention networks for scene graph generation. *Pattern Recognition Letters*, 153:200–206, 2022. 2

[26] 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, 3, 6, 7

[27] Yikang Li, Wanli Ouyang, Xiaogang Wang, and Xiao'ou Tang. Vip-cnn: Visual phrase guided convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1347–1356, 2017. 2

[28] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. Scene graph generation from objects, phrases and region captions. In *Proceedings of the IEEE international conference on computer vision*, pages 1261–1270, 2017. 2, 6

[29] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European Conference on Computer Vision*, 2016. 1

[30] Xinyu Lyu, Lianli Gao, Yuyu Guo, Zhou Zhao, Hao Huang, Heng Tao Shen, and Jingkuan Song. Fine-grained predicates learning for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19467–19475, 2022. 2

[31] RT McAllister, Yarin Gal, Alex Kendall, Mark Van Der Wilk, Amar Shah, Roberto Cipolla, and Adrian Weller. Concrete problems for autonomous vehicle safety: Advantages of bayesian deep learning. *International Joint Conference on Artificial Intelligence*, 2017. 1

[32] Sayak Nag, Kyle Min, Subarna Tripathi, and Amit K Roy-Chowdhury. Unbiased scene graph generation in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22803–22813, 2023. 1, 2

[33] Harris Papadopoulos, Volodya Vovk, and Alex Gammerman. Conformal prediction with neural networks. *19th IEEE International Conference on Tools with Artificial Intelligence*, 2007. 1, 3

[34] Krishan Rana, Jesse Haviland, Sourav Garg, Jad Abou-Chakra, Ian Reid, and Niko Suenderhauf. Sayplan: Grounding large language models using 3d scene graphs for scalable robot task planning. In *7th Annual Conference on Robot Learning*, 2023. 2

[35] Allen Z. Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, Zhenjia Xu, Dorsa Sadigh, Andy Zeng, and Anirudha Majumdar. Robots that ask for help: Uncertainty alignment for large language model planners. In *Proceedings of The 7th Conference on Robot Learning*, pages 661–682. PMLR, 2023. 1, 2

[36] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 3, 6

[37] Yaniv Romano, Rina Foygel Barber, Chiara Sabatti, and Emmanuel J Candès. With Malice Towards None: Assessing Uncertainty via Equalized Coverage. *Harvard Data Science Review*, 2020. 3

[38] Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591, 2020. 2, 3

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

[40] Matteo Sesia and Yaniv Romano. Conformal Prediction using Conditional Histograms. *Advances in Neural Information Processing Systems*, 2021. 3

[41] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 2008. 3

[42] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adri Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*, 2023. 5

[43] Kamile Stankeviciute, Ahmed M. Alaa, and Mihaela van der Schaar. Conformal time-series forecasting. In *Advances in Neural Information Processing Systems*, pages 6216–6228. Curran Associates, Inc., 2021. 3

[44] 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. 6, 7

[45] 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 (CVPR)*, 2020. 3

[46] 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. 2

[47] Leitian Tao, Li Mi, Nannan Li, Xianhang Cheng, Yaosi Hu, and Zhenzhong Chen. Predicate correlation learning for scene graph generation. *IEEE Transactions on Image Processing*, 31:4173–4185, 2022. 2

[48] Damien Teney, Lingqiao Liu, and Anton van Den Hengel. Graph-structured representations for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2017. 1

[49] Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019. 3

[50] Alexander Timans, Christoph-Nikolas Straehle, Kaspar Sakmann, and Eric Nalisnick. Adaptive bounding box uncertainties via two-step conformal prediction, 2024. 1, 2, 3, 6

[51] Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pages 475–490. PMLR, 2012. 6

[52] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer-Verlag, Berlin, Heidelberg, 2005. 3

[53] Vladimir Vovk, A. Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005. 1, 3

[54] Gengcong Yang, Jingyi Zhang, Yong Zhang, Baoyuan Wu, and Yujiu Yang. Probabilistic modeling of semantic ambiguity for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12527–12536, 2021. 1, 2

[55] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation, 2018. 2

[56] Ruichi Yu, Ang Li, Vlad I Morariu, and Larry S Davis. Visual relationship detection with internal and external linguistic knowledge distillation. In *Proceedings of the IEEE international conference on computer vision*, pages 1974–1982, 2017. 2

[57] 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. 3, 6, 7

[58] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Conference on Computer Vision and Pattern Recognition*, 2018. 2

[59] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. Visual translation embedding network for visual relation detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5532–5540, 2017. 2