Conformal Prediction and MLLM aided Uncertainty Quantification in Scene Graph Generation

POSTER #99

Sayak Nag

Udita Ghosh

Calvin-Khang Ta

Sarosij Bose

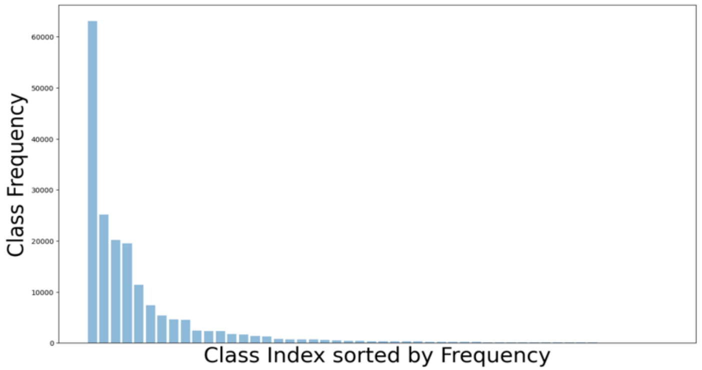
Jiachen Li

Amit K. Roy-Chowdhury

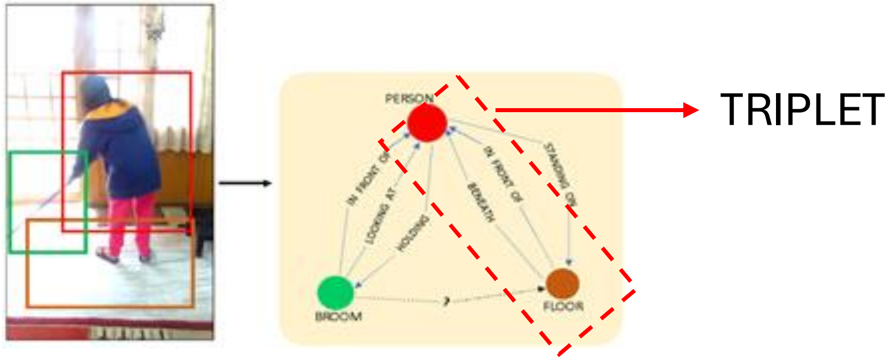
UCR Vision and Learning Group

CVPR Nashville JUNE 11-15, 2025

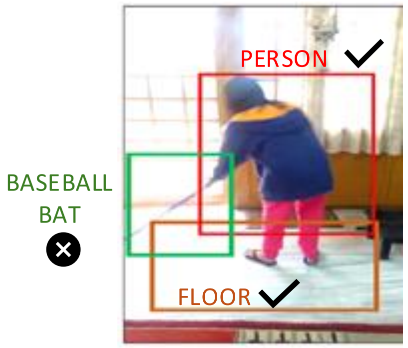
MOTIVATION



Long-tail distribution of Relationships/Predicates



Imprecise or missing scene descriptions



Object detection failure

- Results in generation of noisy scene graphs
- Necessary to quantify the uncertainty of SGG methods in a *post-hoc* manner with statistical guarantees.

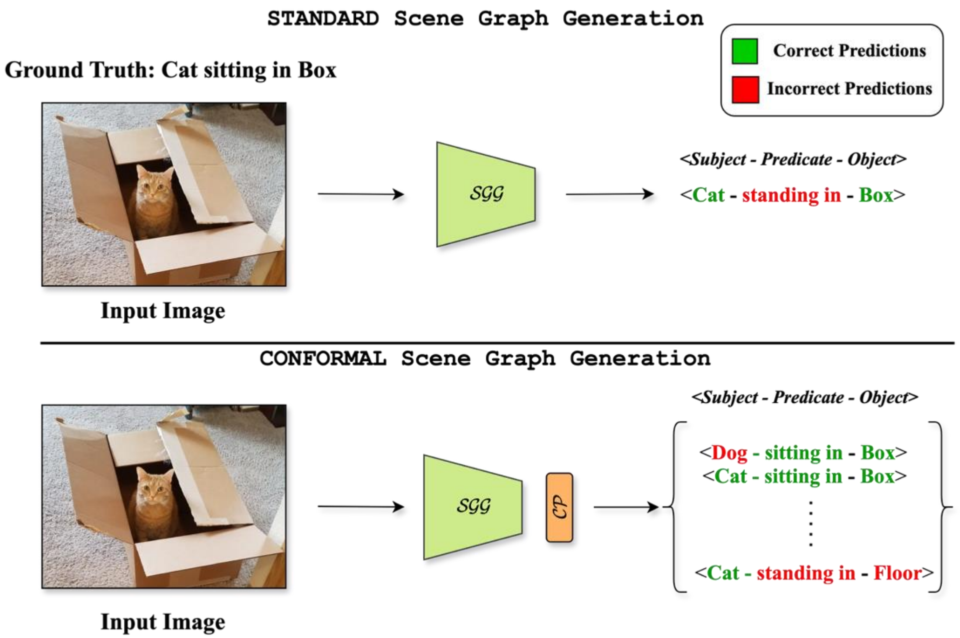
CONFORMAL PREDICTION

OVERVIEW

- Given dataset $\mathcal{D} = \{\mathcal{D}_{tr}, \mathcal{D}_{cal}, \mathcal{D}_{test}\}$
- Under the assumption of of exchangeability $\mathcal{D}_{cal} \cup (X_{n+1}, Y_{n+1})$,
  - $P(Y_{n+1} \in \hat{\mathcal{C}}(X_{n+1})) \geq 1 - \alpha$
- Class-conditional conformal prediction,
  - $P(Y_{n+1} \in \hat{\mathcal{C}}(X_{n+1}) \mid Y_{n+1} = y) \geq 1 - \alpha_y \quad \forall y \in \mathcal{Y}$

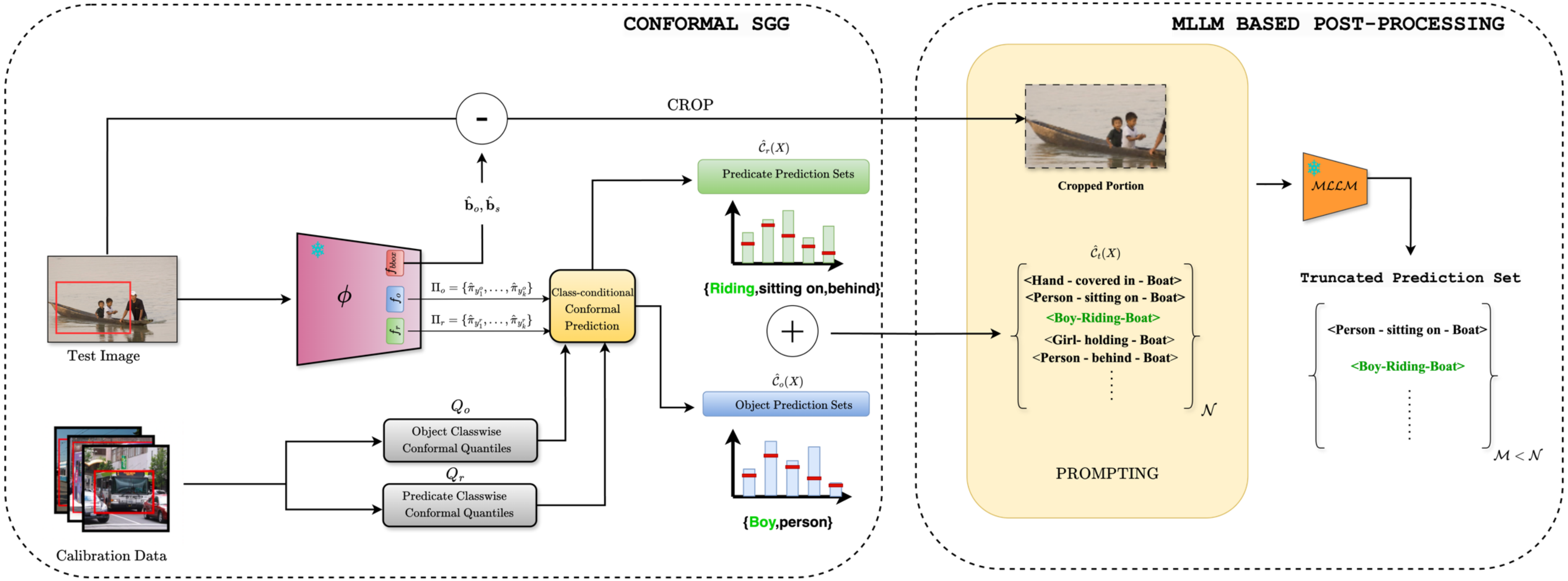
CONFORMAL CALIBRATION

- Design a nonconformity measure $A()$
- Compute class-wise non-conformity scores for samples in $\mathcal{D}_{cal}$
- Sort class-wise non-conformity scores to obtain $1 - \alpha_y$ th quantile.



FRAMEWORK

Plausibility ensured Conformal SGG or PC-SGG



$$\mathcal{A} : \mathcal{X} \times \mathcal{Y} \to [0,1],\ (\hat{f}(X), y) \mapsto 1 - \hat{\pi}_y(X)$$

Non-conformity measure

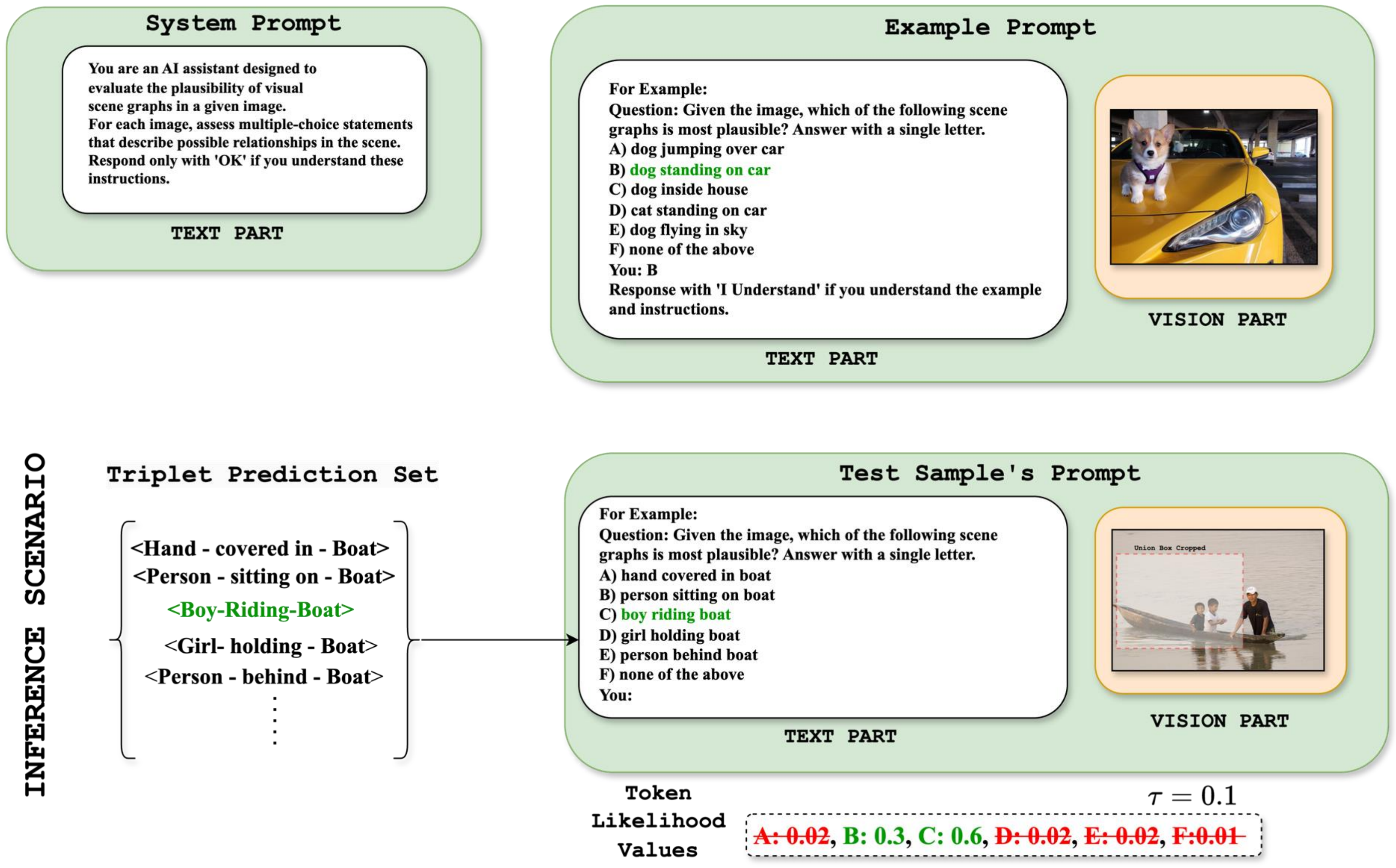
$$\hat{\mathcal{C}}_o(X^o_{n+1}) = \{y^o_k \in \mathcal{Y}_o : \hat{\pi}_{y^o_k} \geq 1 - \hat{q}_{y^o_k}\}$$
$$\hat{\mathcal{C}}_r(X^r_{n+1}) = \{y^r_k \in \mathcal{Y}_r : \hat{\pi}_{y^r_k} \geq 1 - \hat{q}_{y^r_k}\}$$

Object and Predicate Conformal Sets

$$\hat{q}_{y^o_i} = \lceil (n_{y^o_i} + 1)(1 - \alpha_o) / n_{y^o_i} \rceil$$
$$\hat{q}_{y^r_i} = \lceil (n_{y^r_i} + 1)(1 - \alpha_r) / n_{y^r_i} \rceil$$

Class-wise calibrated quantile scores

MLLM BASED POST-PROCESSING



- Leverages one shot in context learning in a MCQA setup.
- Compresses prediction sets into most plausible ones.
- MLLM: BLIP2-Flan-T5-XL

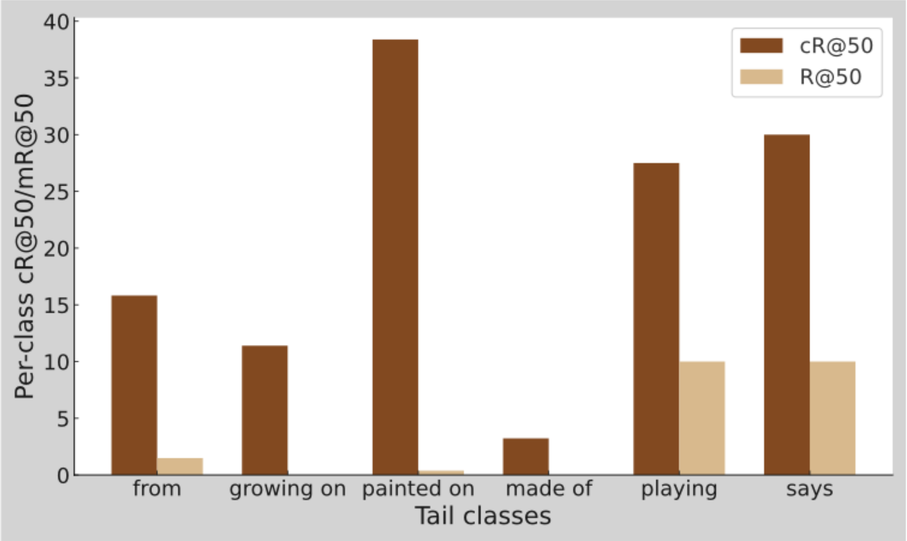
CONFORMAL SGG COVERAGE GUAURANTEE

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 $P(y_k^t \in \hat{\mathcal{C}}_t(X_{n+1}^r)) = P(Y_{n+1}^o \in \hat{\mathcal{C}}_o(X_{n+1}^o) \mid Y_{n+1}^o = y_i^o) \cdot P(Y_{n+1}^r \in \hat{\mathcal{C}}_o(X_{n+1}^r) \mid Y_{n+1}^r = y_m^r)\ \forall\, y_k^s \in \mathcal{Y}_o, y_k^o \in \mathcal{Y}_o, y_k^r \in \mathcal{Y}_r$.*

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

RESULTS

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**



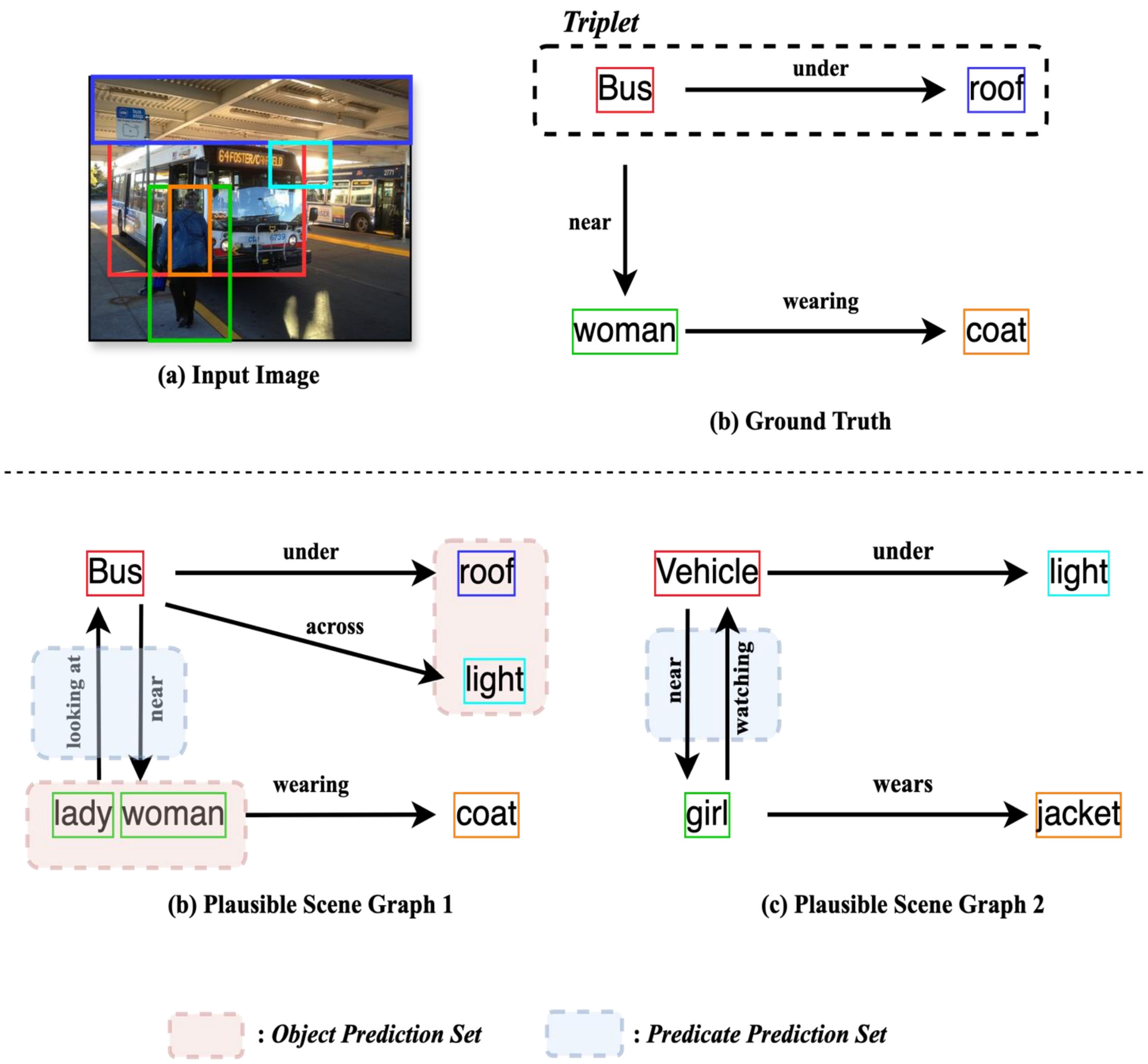
Method	w/o MLLM Plausibility Assessment		w/ MLLM Plausibility Assessment	
	$Cov_T \uparrow$	$AvgSize \downarrow$	$Cov_T \uparrow$	$AvgSize \downarrow$
MOTIFS [57]	74.97	866.09	74.97	403.21
MOTIFS-D [11]	76.93	893.21	76.67	411.58
VCTREE [44]	80.06	818.76	80.06	389.24
SQUAT [20]	80.43	816.68	80.25	398.67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

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**

$$\alpha_o = 0.05\,, \alpha_r = 0.1$$

RESULTS


Triplet
Bus
under
roof
near
woman
wearing
coat
(a) Input Image
(b) Ground Truth
Bus
under
roof
across
light
looking at
near
lady
woman
wearing
coat
Vehicle
under
light
near
watching
girl
wears
jacket
(b) Plausible Scene Graph 1
(c) Plausible Scene Graph 2
: Object Prediction Set
: Predicate Prediction Set

Poster Session: June 14, Exhall D
Poster ID: 99

PAPER