AAAI-25

FEBRUARY 25 – MARCH 4, 2025 | PHILADELPHIA, USA

RA-SGG: Retrieval-Augmented Scene Graph Generation Framework via Multi-Prototype Learning

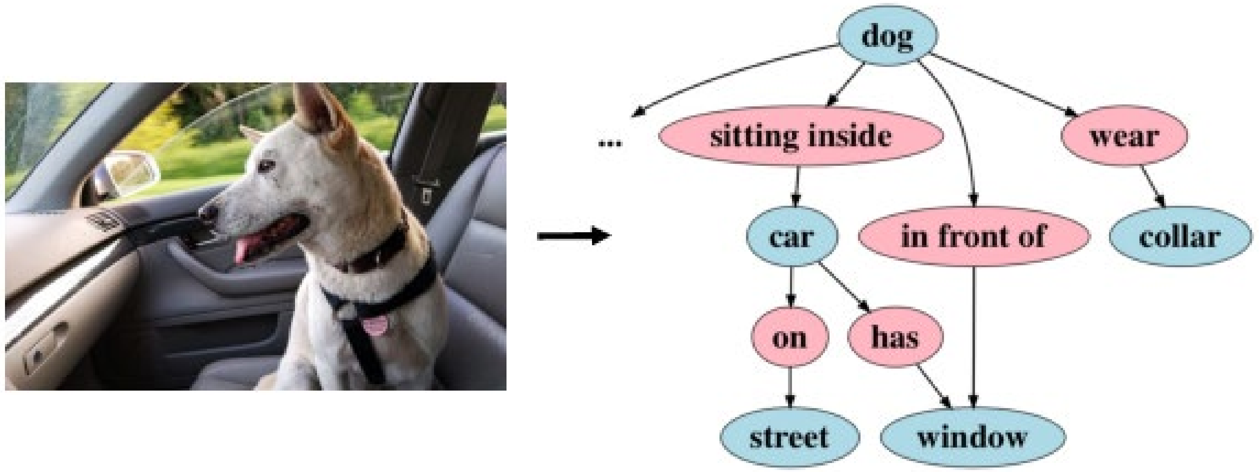
Kanghoon Yoon, Kibum Kim, Yeonjun In,
Jaehyeong Jeon, Donghyun Kim, Chanyoung Park

Department of Industrial & Systems Engineering
KAIST
ykhoon08@kaist.ac.kr

Problem Statement

Scene Graph Generation

- Scene Graph Generation (SGG) aims to detecting objects within images and predicting relationships between them.
  - Each object including bounding box information represents a node of the scene graph
  - Each predicates represents an edge of the scene graph

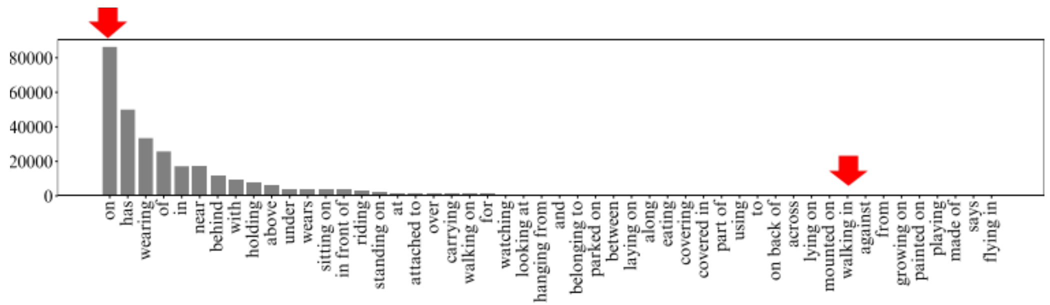


Problem Statement

Scene Graph Generation

- Scene Graph Generation (SGG) faces two major challenges:

Long-tailed Distribution



It leads to training a SGG model that predicts majority classes

Semantic Ambiguity

on
stand on
walk on

carrying
holding
playing

Model become confused with semantically ambiguous predicates

These lead to bias towards head predicates and poor fine-grained relationship detection!

Limitation of Existing SGG Works

- Existing SGG works rely on single-label classification formulation of the problem

GT: < cat, on, table >

<cat, on, table> ?

<cat, lying on, table> ?

<cat, laying on, table> ?

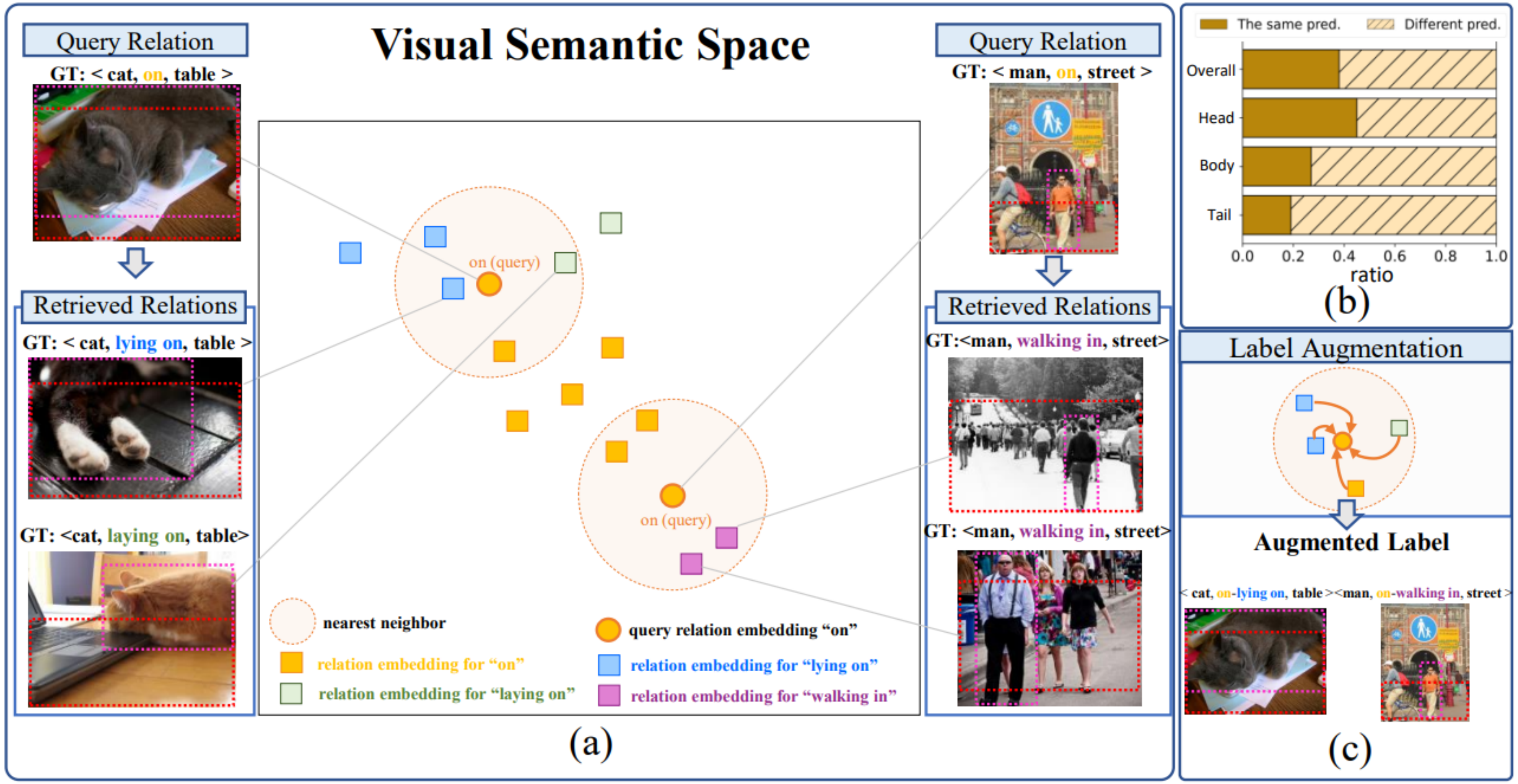
- However,
  - it forces model to select just one predicate while suppressing others due to the semantic ambiguity.
  - Ignores the nature of natural language where multiple predicates can describe same relationship.

We argue that addressing the long-tailed problem and semantic ambiguity is difficult under the single-label classification formulation of SGG problem.

Main Idea of Our Work (RA-SGG)

RA-SGG reframes SGG as multi-label classification with partial annotation

- **Utilize semantically similar predicates in the visual semantic embedding space!**



- Identify potentially multi-labeled instances and augment the predicate labels

Our Formulation of the SGG Problem

We reframe SGG as multi-label classification with partial annotation

- Assume that we have only partial (single) annotations among multiple annotations.
  - i.e., the predicates from true unbiased data distribution is $y_i^* \in \{0,1\}^{N_p}$ $(\sum_i y_i \geq 1)$.
  - However, we only have the predicates of observed samples $y_i \in \{0,1\}^{N_p}$ $(\sum_i y_i = 1)$

- To obtain the unbiased model, we can minimize the following estimated loss called inverse propensity scored loss as follows:

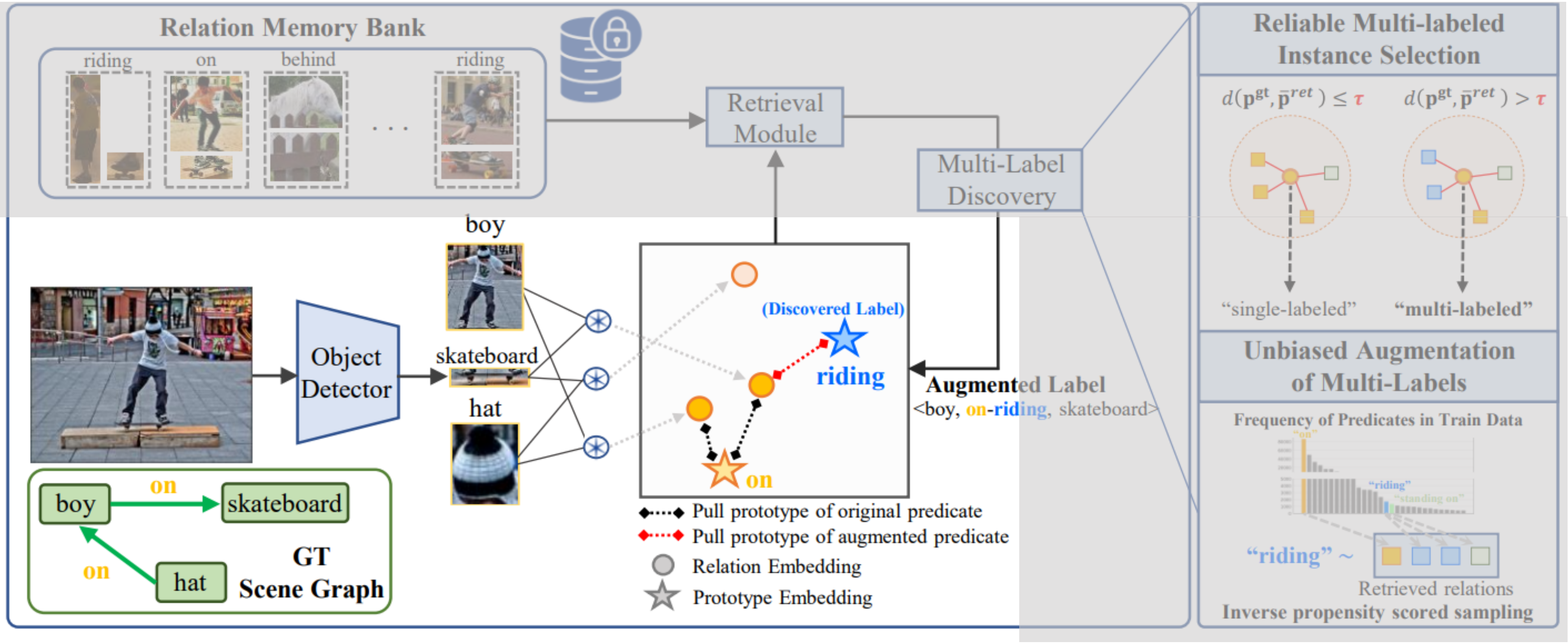
$$\mathcal{L}_{ips} = -\sum_{i=1}^{N_p} \underbrace{P(\mathbf{y}_i = 1 | \mathbf{y}_i^* = 1)^{-1}}_{\text{inverse propensity score}} \mathbf{y}_i \log \hat{\mathbf{y}}_i$$

  - Instead of directly minimizing this inverse propensity-scored loss, we will estimate this loss through retrieval-augmented framework.
    - We estimate the loss by finding and augmenting more samples based on inverse propensity.

Pipeline of RA-SGG

Retrieval Augmented Scene Graph Generation Framework

Phase 1. Train Prototype Embedding Network using GT Scene Graph

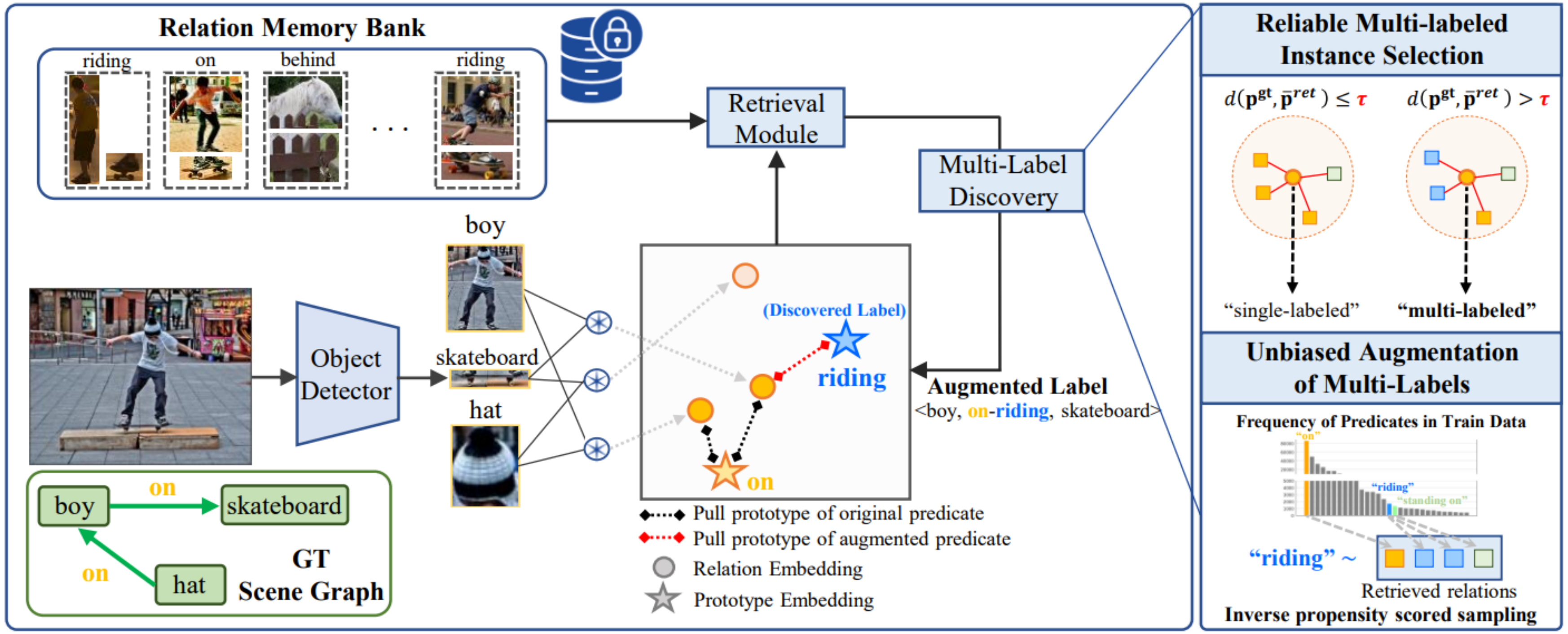


- Generates relation features through fusion layer, which is applied to the subject-object features.
- Minimize distance between relation features with their ground truth prototype

Pipeline of RA-SGG

Retrieval Augmented Scene Graph Generation Framework

Phase 2. Train Prototype Embedding Network using GT Scene Graph and Augmented Label



- Find semantically similar instances from the established relation memory bank.
- Minimize distance between relation features with their ground truth prototype and the prototype of the augmented label

RA-SGG

Relation memory bank

- Memory bank includes key-value pair, which consists of the relation embedding $\mathbf{r}$ and its GT predicate $\mathbf{p}$
  - **i.e., memory bank =** $\{(\mathbf{r_1}, \mathbf{p_1}), (\mathbf{r_2}, \mathbf{p_2}), \dots, (\mathbf{r_i}, \mathbf{p_i}), \dots, (\mathbf{r_M}, \mathbf{p_M})\}$

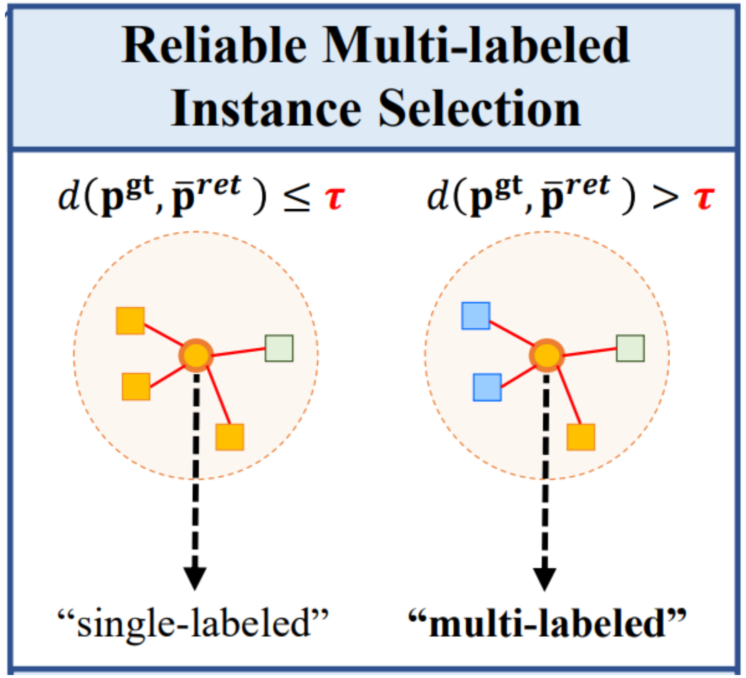
- Given an image, we obtains the relation embedding $\boldsymbol{r}$ between subject and object features using SGG models like PE-Net.

- Retrieve the top-K relevant relation instances from the memory bank using cosine similarity between relation embeddings
  - Given query embedding $\mathbf{r}$, obtain $(\mathbf{r}_1^{ret}, \mathbf{p}_1^{ret}), (\mathbf{r}_2^{ret}, \mathbf{p}_2^{ret}), \dots, (\mathbf{r}_K^{ret}, \mathbf{p}_K^{ret})$

RA-SGG

How can we obtain reliable multi-labeled instances?

- Given retrieved instances $(\mathbf{r}_1^{ret}, \mathbf{p}_1^{ret}), (\mathbf{r}_2^{ret}, \mathbf{p}_2^{ret}), \dots, (\mathbf{r}_K^{ret}, \mathbf{p}_K^{ret})$, we use label inconsistency score to identify potential multi-label instances.
- Label Inconsistency Score computes the Euclidean distance $d(\cdot)$ between $\mathbf{p}^{gt}$ and $\overline{\mathbf{p}}^{ret}$.
  - It measures discrepancy between ground-truth and averaged retrieved predicates $\overline{\mathbf{p}}^{ret}$
  - It helps maintain reliability of pseudo-labels



- We finally define single-labeled instance and multi-labeled instance as follows:

$$\mathcal{D}_{\text{single}} \leftarrow \{(\mathbf{s}_i, \mathbf{p}_i, \mathbf{o}_i) | d(\mathbf{p}^{\text{q}}, \bar{\mathbf{p}}^{\text{ret}}) < \tau, \forall (\mathbf{s}_i, \mathbf{p}_i, \mathbf{o}_i) \in \mathcal{D}_{\text{Tr}}\}$$
$$\mathcal{D}_{\text{multi-}} \leftarrow \{(\mathbf{s}_i, \mathbf{p}_i, \mathbf{o}_i) | d(\mathbf{p}^{\text{q}}, \bar{\mathbf{p}}^{\text{ret}}) \ge \tau, \forall (\mathbf{s}_i, \mathbf{p}_i, \mathbf{o}_i) \in \mathcal{D}_{\text{Tr}}\}$$

$$\mathcal{L}_{ips} = -\sum_{i=1}^{N_p} \underbrace{P(\mathbf{y}_i = 1|\mathbf{y}_i^* = 1)^{-1}}_{\text{inverse propensity score}} \mathbf{y}_i \log \hat{\mathbf{y}}_i$$

RA-SGG

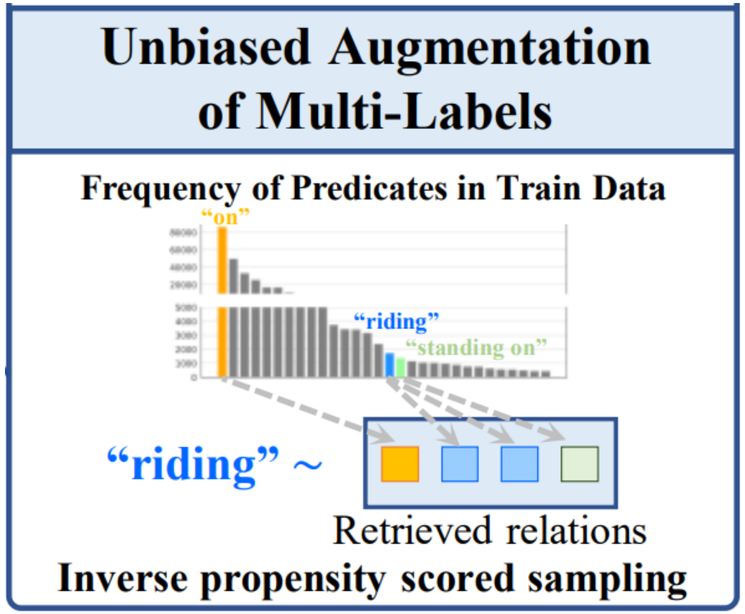
How can we select the augmented predicates?

- We compute averaged inverse propensity of retrieved instances
  - The propensity of each predicate is the frequency in the training data
  - The averaged inverse propensity of retrieved instances

$$w = \text{Softmax}(\textstyle\sum_{k=1}^{K} s_k^{\text{ret}} \mathbf{p}_k^{\text{ret}})$$

  - This inverse propensity encourage RA-SGG to sample tail predicates rather than head predicates.

  - Note that some predicates such as includes extremely small number of samples in the training data
    - E.g., "flying in" includes less than 10 samples in the training data.



- We argue that inverse propensity-based augmentation strategy is more effective compared to minimizing L_{ips}.

Experiment

Experimental settings and datasets

Dataset

- Visual Genome (150 objects, 50 predicates)
- GQA (200 objects, 100 predicates)

Evaluation Protocol

- Predcls: Predict predicate class given GT object bounding boxes, and their GT object classes are given
- SGCls: Predict predicate class and object classes given GT object bounding boxes
- SGDet: Predict predicate classes, object classes, and bounding boxes

Backbone: ResNeXt-101-FPN with Faster R-CNN

Metric: Recall@K, meanRecall@K, and Harmonic mean of previous two metrics (F@K)

Experiment

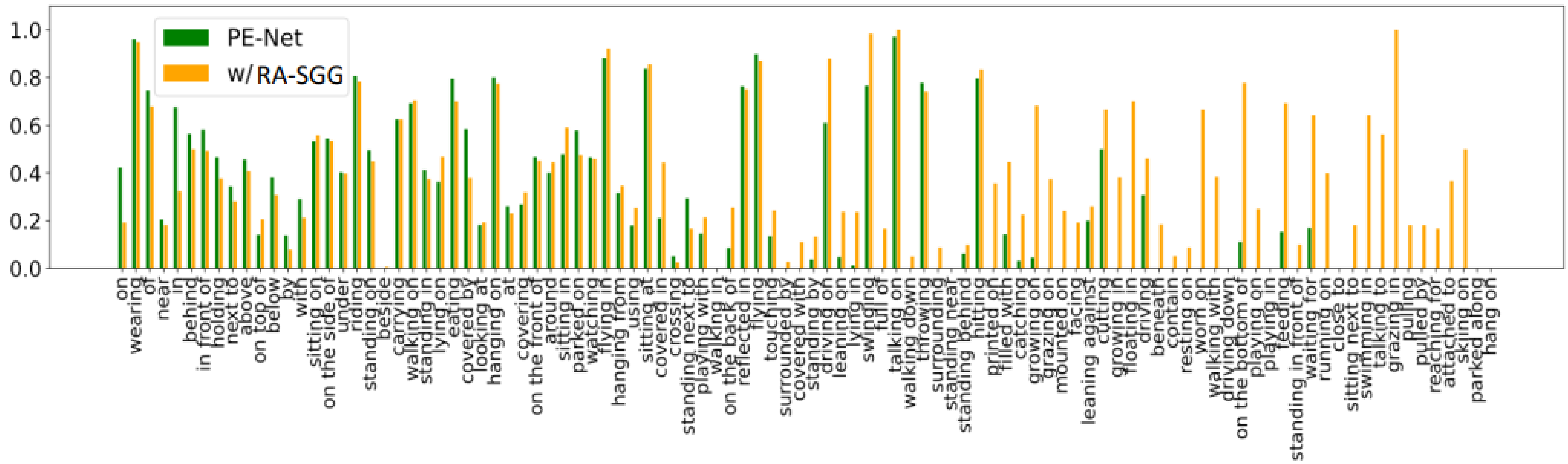
Result on Visual Genome Dataset

B	Methods	Predicate Classification			Scene Graph Classification			Scene Graph Detection		
		R@50/100	mR@50/100	F@50/100	R@50/100	mR@50/100	F@50/100	R@50/100	mR@50/100	F@50/100
Specific	KERN(Chen et al. 2019)CVPR'19	65.8/67.6	17.7/19.2	27.9/29.9	36.7/37.4	9.4/10.0	15.0/15.8	27.1/29.8	6.4/7.3	10.4/11.7
	BGNN(Li et al. 2021)CVPR'21	59.2/61.3	30.4/32.9	40.2/42.8	37.4/38.5	14.3/16.5	20.7/23.1	31.0/35.8	10.7/12.6	15.9/18.6
	DT2ACBS(Desai et al. 2021)ICCV'21	23.3/25.6	35.9/**39.7**	28.3/31.1	16.2/17.6	**24.8/27.5**	19.6/21.5	15.0/16.3	**22.0/24.0**	17.8/19.4
	HL-Net(Lin et al. 2022)CVPR'22	67.0/68.9	- /22.8	- /34.3	42.6/43.5	- /13.5	- /20.6	33.7/38.1	- /9.2	- /14.8
	HetSGG(Yoon et al. 2023)AAAI'23	57.8/59.1	31.6/33.5	40.9/42.8	37.6/38.7	17.2/18.7	23.6/25.2	30.0/34.6	12.2/14.4	17.3/20.3
	SQUAT(Jung et al. 2023)CVPR'23	55.7/57.9	30.9/33.4	39.7/42.4	33.1/34.4	17.5/18.8	22.9/24.3	24.5/28.9	14.1/16.5	17.9/21.0
Motif	Motif(Zellers et al. 2018)CVPR'18	64.6/66.0	15.2/16.2	24.6/26.0	38.0/38.9	8.7/9.3	14.2/15.0	31.0/35.1	6.7/7.7	11.0/12.6
	TDE(Kaihua et al. 2020)CVPR'20	46.2/51.4	25.5/29.1	32.9/37.2	27.7/29.9	13.1/14.9	17.8/19.9	16.9/20.3	8.2/9.8	11.0/13.2
	DLFE(Chiou et al. 2021)MM'21	52.5/54.2	26.9/28.8	35.6/37.6	32.3/33.1	15.2/15.9	20.7/21.5	25.4/29.4	11.7/13.8	16.0/18.8
	NICE(Li et al. 2022)CVPR'22	55.1/57.2	29.9/32.3	38.8/41.3	33.1/34.0	16.6/17.9	22.1/23.5	27.8/31.8	12.2/14.4	17.0/19.8
	GCL(Dong et al. 2022)CVPR'22	42.7/44.4	36.1/38.2	39.1/41.1	26.1/27.1	20.8/21.8	23.2/24.1	18.4/22.0	16.8/19.3	17.6/20.6
	IETrans(Zhang et al. 2022)ECCV'22	54.7/56.7	30.9/33.6	39.5/42.2	32.5/33.4	16.8/17.9	22.2/23.3	26.4/30.6	12.4/14.9	16.9/20.0
	CFA (Li et al. 2023)ICCV'23	54.1/56.6	35.7/38.2	43.0/45.6	34.9/36.1	17.0/18.4	22.9/24.4	27.4/31.8	13.2/15.5	17.8/20.8
	ST-SGG(Kim et al. 2024a)ICLR'24	53.9/57.7	28.1/31.5	36.9/40.8	33.4/34.9	16.9/18.0	22.4/23.8	26.7/30.7	11.6/14.2	16.2/19.4
PE-Net	PE-Net†(Zheng et al. 2023)CVPR'23	64.9/67.2	31.5/33.8	42.4/45.0	39.4/40.7	17.8/18.9	24.5/25.8	30.7/35.2	12.4/14.5	17.7/20.4
	IETrans†(Zhang et al. 2022)ECCV'22	49.3/51.8	33.5/36.0	39.9/42.5	31.2/32.3	18.3/19.4	23.1/24.2	24.2/28.4	13.7/16.2	17.5/20.6
	CFA†(Li et al. 2023)ICCV'23	57.8/61.6	30.0/33.2	39.5/43.1	36.2/37.1	15.9/18.2	22.1/24.4	25.6/29.8	14.4/17.1	18.4/21.7
	RA-SGG	62.2/64.1	**36.2**/39.1	**45.7/48.6**	38.2/39.1	20.9/22.5	**27.0/28.6**	26.0/30.3	14.4/17.1	**18.5/21.9**

Table 1: Performance (%) of state-of-the-art SGG models on Visual Genome (Krishna et al. 2017). F@K is the harmonic mean of mR@50/100 and R@50/100. † denotes the result produced by us using their official code.

Experiment

Result on GQA 200 Dataset

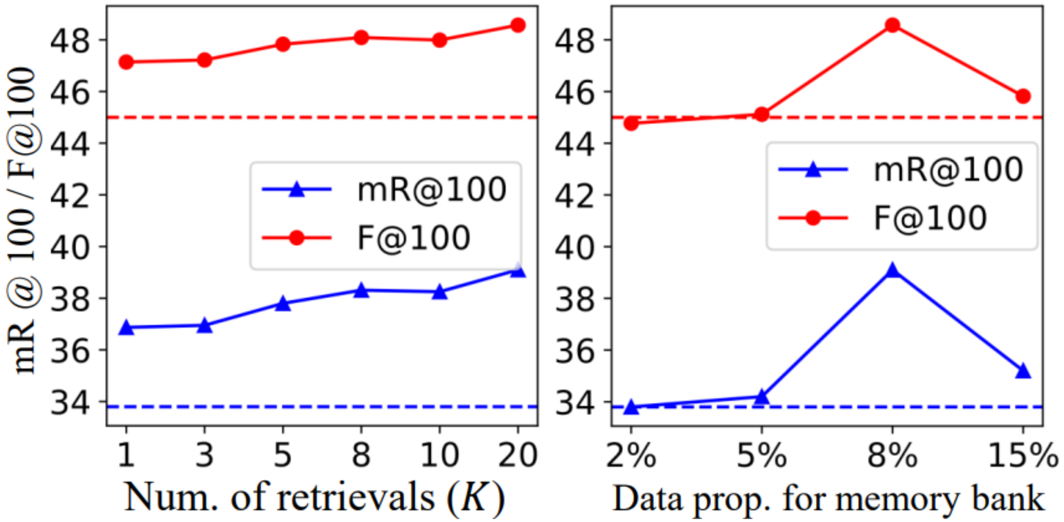


Experiment

Ablation Study of RA-SGG

Model	Predicate Classification			Scene Graph Classification		
	R@50/100	mR@50/100	F@50/100	R@50/100	mR@50/100	F@50/100
Vanilla PE-Net	64.9/67.2	31.5/33.8	42.4/45.0	39.4/40.7	17.8/18.9	24.5/25.8
RA-SGG w/o select.	64.4/66.4	33.4/36.4	44.0/47.0	38.5/39.4	19.6/20.9	26.0/27.3
RA-SGG w/o IPSS	64.6/66.7	32.9/35.1	43.6/46.0	38.6/39.5	18.6/19.8	25.1/26.3
RA-SGG	62.2/64.1	**36.2/39.1**	**45.7/48.6**	38.2/39.1	**20.9/22.5**	**27.0/28.6**

Table 3: Ablation study of RA-SGG.



Conclusion

- The paper reformulates Scene Graph Generation (SGG) as a multi-label classification problem with partial annotation, offering a novel perspective that aligns with natural language's ability to describe the same relationship in multiple ways.

- RA-SGG successfully addresses core SGG challenges by using retrieval-augmentation to discover latent fine-grained predicates while maintaining performance on general predicates, avoiding the common trade-off in previous approaches.

- The framework demonstrates substantial improvements across all predicate types by leveraging retrieval-based multi-label discovery, showing the effectiveness of considering multiple valid predicate descriptions
for a single relationship

AAAI-25

Thank you for listening!