NEURAL INFORMATION PROCESSING SYSTEMS

Interaction-Centric Knowledge Infusion and Transfer for Open Vocabulary Scene Graph Generation

Lin Li[1,2], Chuhan Zhang[1,2], Dong Zhang[1,2], Chong Sun[3], Chen Li[3], Long Chen[1†]

[1]HKUST [2]ACCESS [3]Tencent

1 Definition of OVSGG

NEURAL INFORMATION PROCESSING SYSTEMS

Definition:

Given an image I, Scene Graph Generation (SGG) aims to construct a structured semantic graph $G = (V, E)$. Each node $v_i \in V$ is defined by its bounding box (bbox) and category, while each edge $e_{ij} \in E$ represents the relationship between v_i and v_j.

In **open-vocabulary settings**, the label set C for nodes and edges is divided into base classes C_B and novel classes C_N , such that $C_B \cup C_N = C$ and $C_B \cap C_N = \emptyset$. C_B contains **seen** classes during training, while C_N includes **unseen** classes that the model is expected to generalize to during inference.

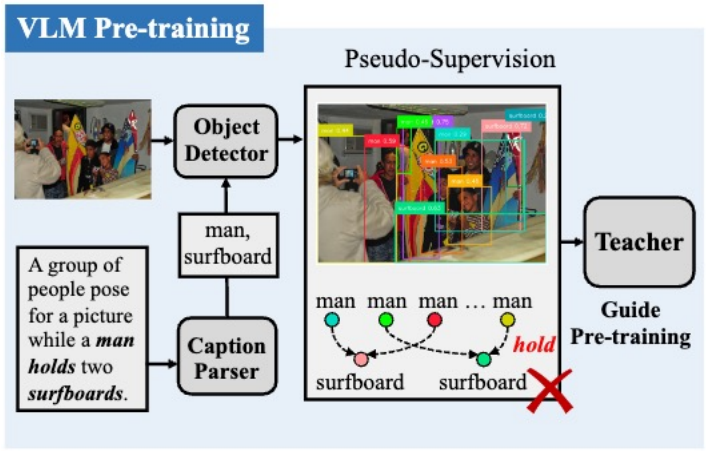
Motivation

NEURAL INFORMATION PROCESSING SYSTEMS

Challenge

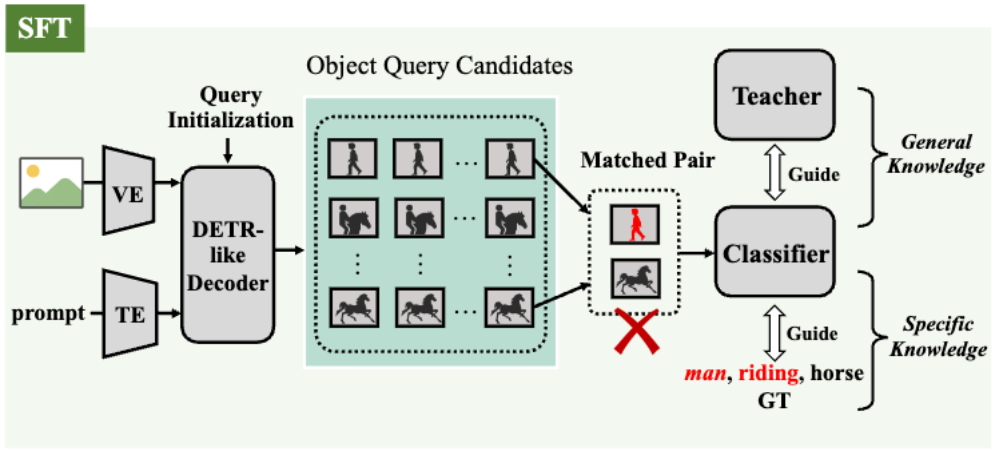
➢ Knowledge Infusion

Using solely object categories for detection causes ambiguity in associating object pairs



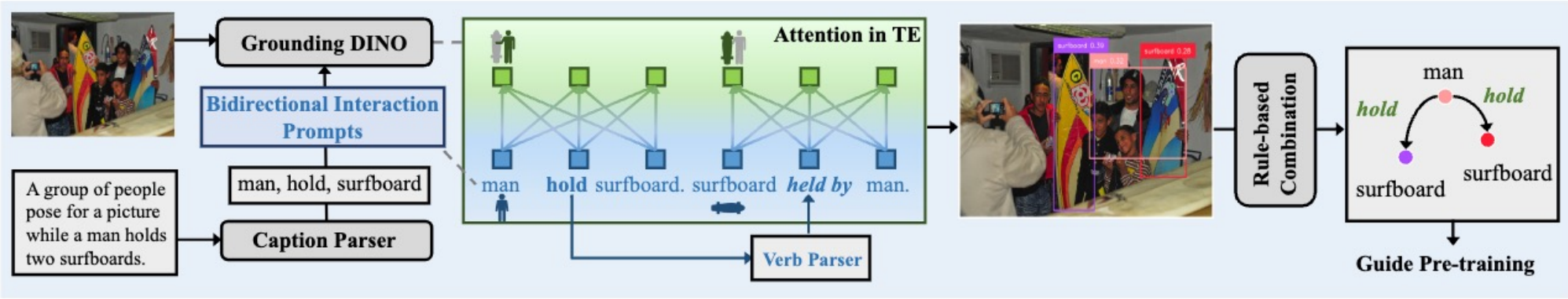
➢ Knowledge Transfer

Vast object query candidates make misaligned non-interacting objects with interacting training target



3 Method

Interaction-centric end-to-end OVSGG framework



3 Method

Interaction-Centric Knowledge Infusion

- **Bidirectional interaction prompt** is designed to guide the object localization.
- Combining two perspectives for each interaction triplet: one from the subject's viewpoint and another from the object's perspective via **counter-action prompts**.

Benefits:

- Modeling Context Information
- Enhancing Object Role Awareness

Question: Given the action 'ride', please generate its corresponding counter-action.
Answer: 'be ridden by'.
Question: Given the action 'eat', please generate its corresponding counter-action.
Answer: 'be eaten by'.

Question: Given the action '{**relation**}', please generate its corresponding counter-action.
Answer:

Counter-action generation prompts

Method

Interaction-Centric Knowledge Transfer

Interaction-Guided Query Selection instills an interaction prior into the two-step query generation process to reduce non-interacting candidates.

- Step I identifies the most relevant visual tokens likely to participate in object interactions.

$$s_i = \left(\max(\mathbf{v}_i \mathbf{T}_o^\top)\right)^\gamma \cdot \left(\max(\mathbf{v}_i \mathbf{T}_r^\top)\right)^{1-\gamma},$$

- Step II models interaction semantics by integrating relational context into object tokens.

$$s_i^{in} = \max(\mathbf{v}_i \mathbf{T}_{in}^\top).$$

Method

Interaction-Centric Knowledge Transfer

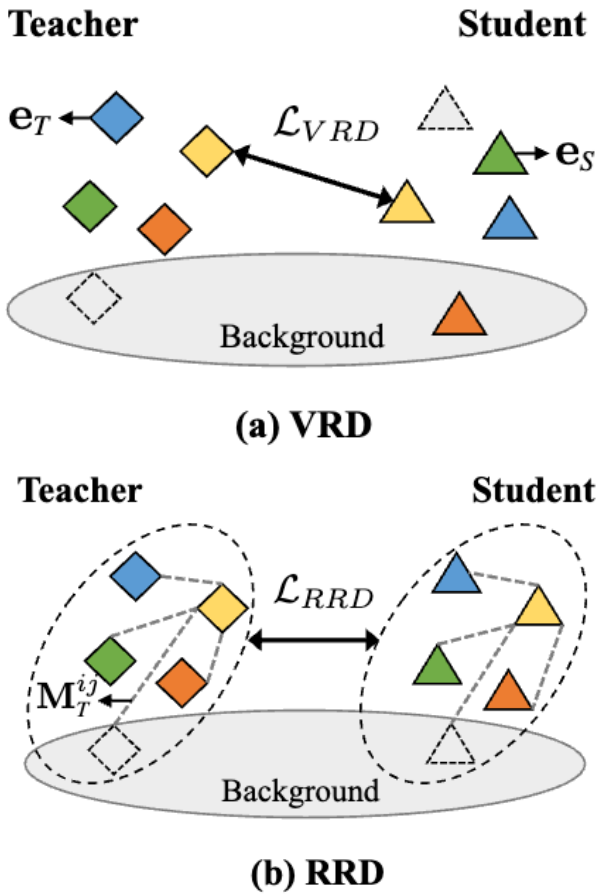
Interaction-Consistent Knowledge Distillation

- **Visual-concept retention distillation**: ensures that the student's edge features remain point-wise consistent with the teacher's semantic space for negative samples.

$$\mathcal{L}_{VRD} = \frac{1}{|\mathcal{N}|} \sum_{\mathbf{e} \in \mathcal{N}} \|\mathbf{e}_S - \mathbf{e}_T\|_1,$$

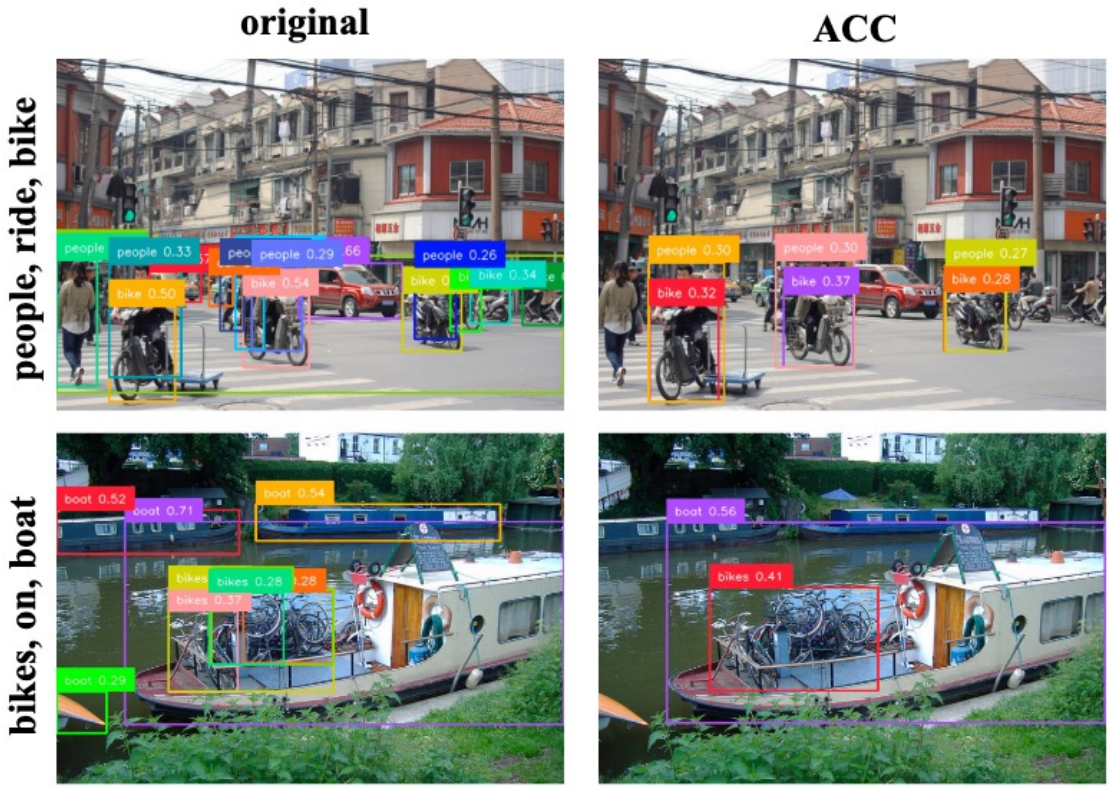
- **Relative-interaction retention distillation**: explicitly models inter-pair relativity by aligning the structure similarity of triplet embeddings between the teacher and student models.

$$\mathbf{M}_T^{ij} = \frac{\mathbf{e}_T^i \cdot \mathbf{e}_T^{j\top}}{\|\mathbf{e}_T^i \cdot \mathbf{e}_T^{j\top}\|_2}, \quad \mathbf{M}_S^{ij} = \frac{\mathbf{e}_S^i \cdot \mathbf{e}_S^{j\top}}{\|\mathbf{e}_S^i \cdot \mathbf{e}_S^{j\top}\|_2}. \quad \mathcal{L}_{RRD} = \frac{1}{|\mathcal{N}|^2} \|\mathbf{M}_S - \mathbf{M}_T\|_F^2.$$

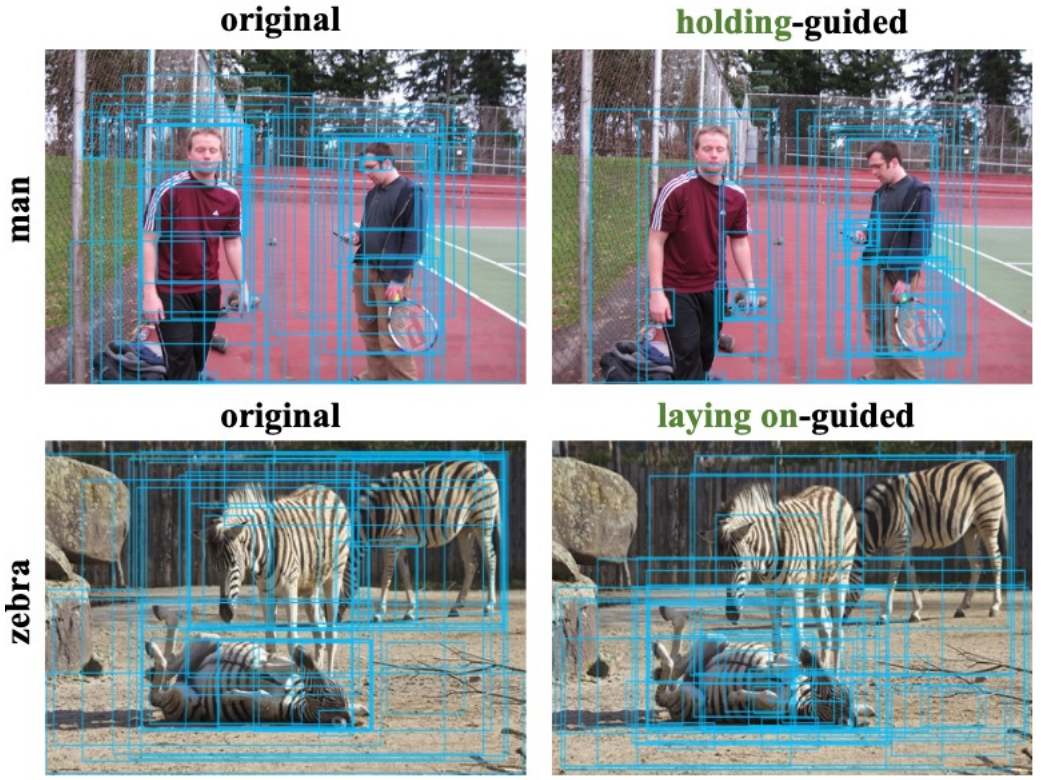


4 Results

NEURAL INFORMATION PROCESSING SYSTEMS



Pseudo supervision generation



Interaction-guided query selection

4 Results

Method		Backbone	Base+Novel (Relation)			Novel (Relation)		
			R@20	R@50	R@100	R@20	R@50	R@100
IMP [55]	CVPR'17	-	-	12.56	14.65	-	0.00	0.00
MOTIFS [61]	CVPR'18	-	-	15.41	16.96	-	0.00	0.00
VCTREE [47]	CVPR'19	-	-	15.61	17.26	-	0.00	0.00
TDE [46]	CVPR'20	-	-	15.50	17.37	-	0.00	0.00
OpenSGen [18]	ICMR'25	-	-	18.00	20.50	-	15.70	17.90
VS^3 [63]	CVPR'23	Swin-T	-	15.60	17.30	-	0.00	0.00
OvSGTR [9]	ECCV'24	Swin-T	-	20.46	23.86	-	13.45	16.19
RAHP [36]	AAAI'25	Swin-T	-	20.50	25.74	-	15.59	19.92
ACC (Ours)		Swin-T	**17.49**	**23.22**	**27.40**	**12.90**	**17.89**	**21.70**
OvSGTR [9]	ECCV'24	Swin-B	-	22.89	26.65	-	16.39	19.72
ACC (Ours)		Swin-B	**18.77**	**24.81**	**29.28**	**14.72**	**20.04**	**24.66**

Experimental results of OvR-SGG setting on VG test set.

Method		Backbone	Joint Base+Novel			Novel (Obj)			Novel (Rel)		
			R@20	R@50	R@100	R@20	R@50	R@100	R@20	R@50	R@100
IMP [55]	CVPR'17	-	-	0.77	0.94	-	0.00	0.00	-	0.00	0.00
MOTIFS [61]	CVPR'18	-	-	1.00	1.12	-	0.00	0.00	-	0.00	0.00
VCTREE [47]	CVPR'19	-	-	1.04	1.17	-	0.00	0.00	-	0.00	0.00
TDE [46]	CVPR'20	-	-	1.00	1.15	-	0.00	0.00	-	0.00	0.00
VS^3 [63]	CVPR'23	Swin-T	-	5.88	7.20	-	0.00	0.00	-	0.00	0.00
OvSGTR [9]	ECCV'24	Swin-T	10.02	13.50	16.37	10.56	14.32	17.48	7.09	9.19	11.18
ACC (Ours)		Swin-T	**12.61**	**17.43**	**21.27**	**12.48**	**17.16**	**21.10**	**11.38**	**15.90**	**19.46**
OvSGTR [9]	ECCV'24	Swin-B	12.37	17.14	21.03	12.63	17.58	21.70	10.56	14.62	18.22
ACC (Ours)		Swin-B	**13.50**	**18.88**	**23.19**	**13.46**	**18.84**	**23.29**	**12.37**	**17.50**	**21.73**

Experimental results of OvD+R-SGG setting on VG test set.

Our code is available at
https://github.com/HKUST-LongGroup/ACC

ACCEPT MY ENDLESS GRATITUDE