Relation Boosting for Unbiased PSG with Vision-Language Model

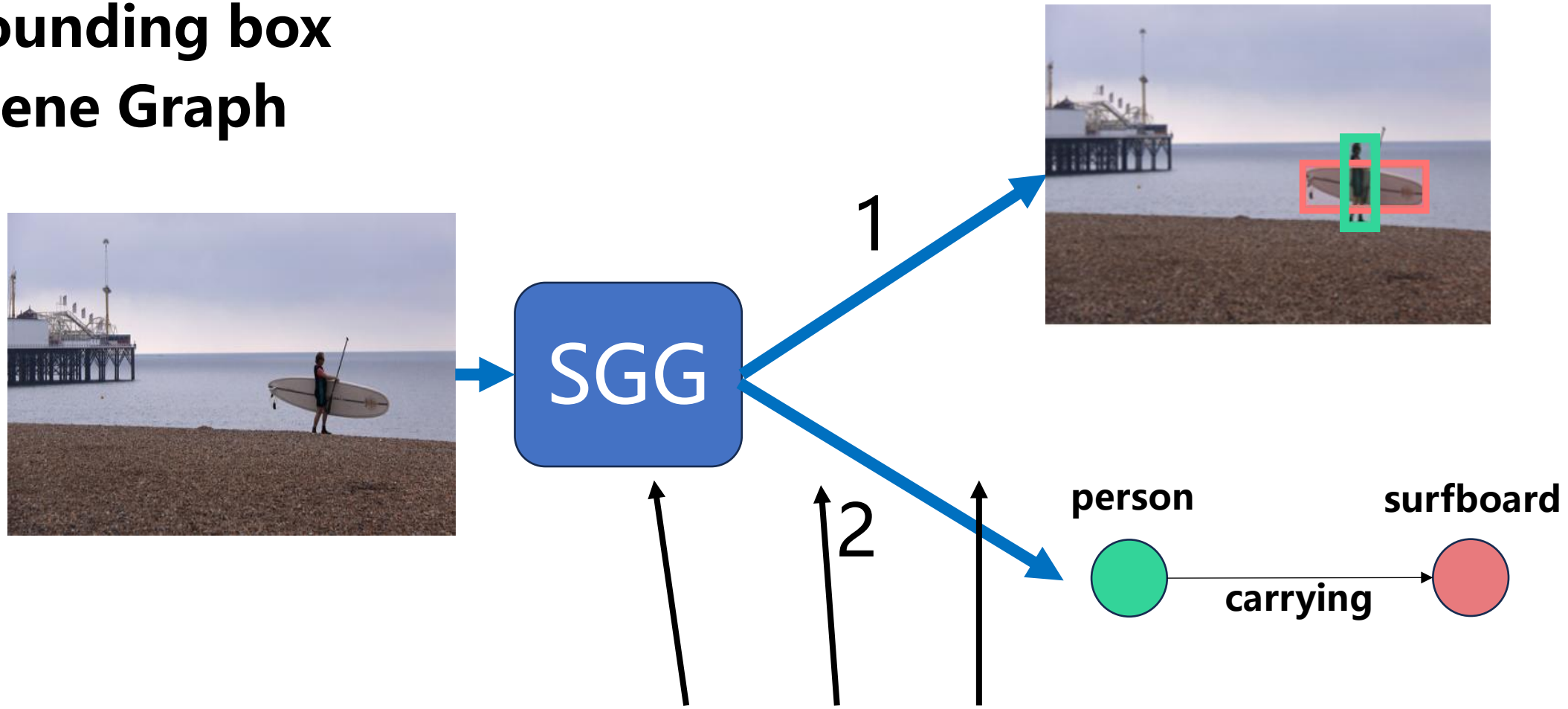
Jingbin Xu , Junwen Chen , Keiji Yanai

The University of Electro-Communications



Scene Graph Generation (SGG)

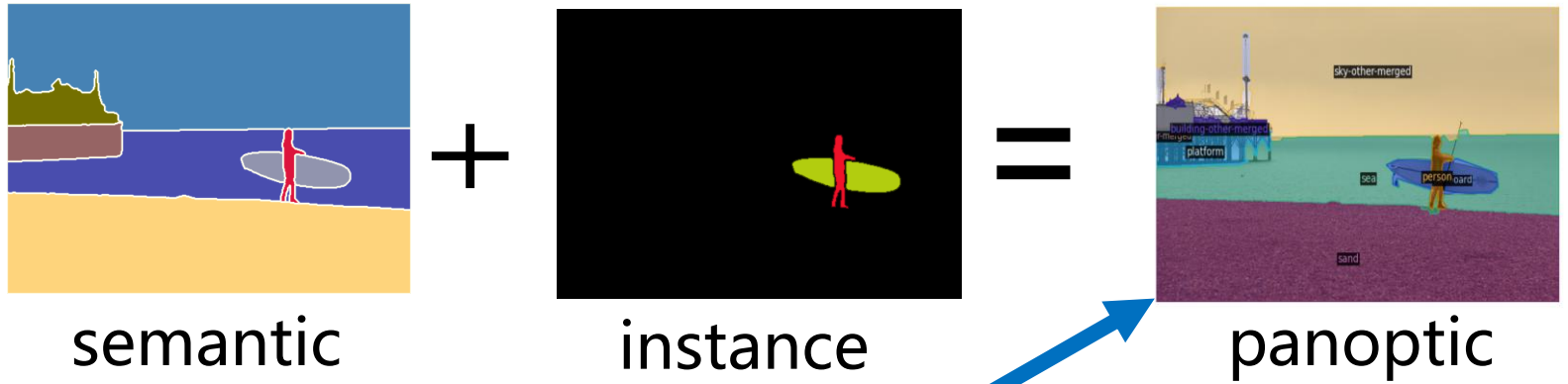
1. Bounding box
2. Scene Graph

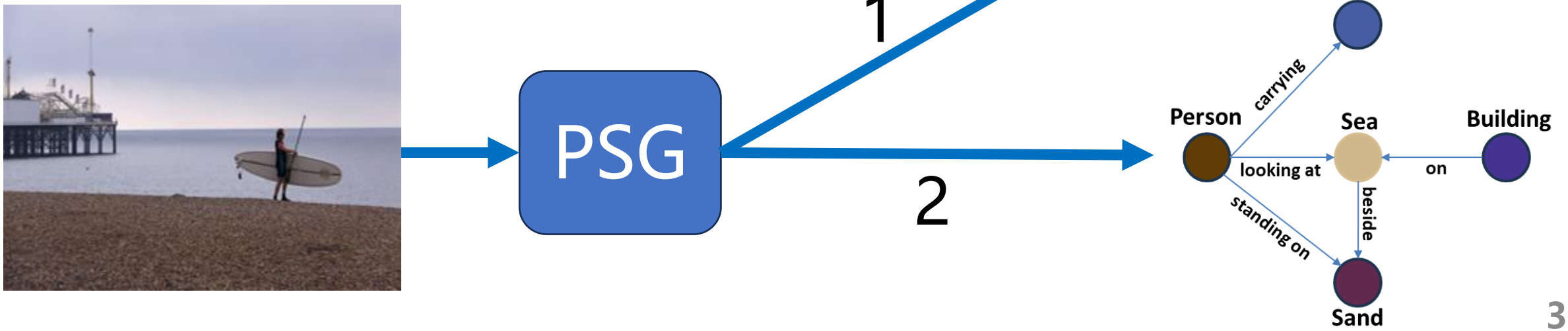


Combine scene graphs by detecting triples of < subject, relation,object > from a single image.

Panoptic Scene Graph Generation (PSG)

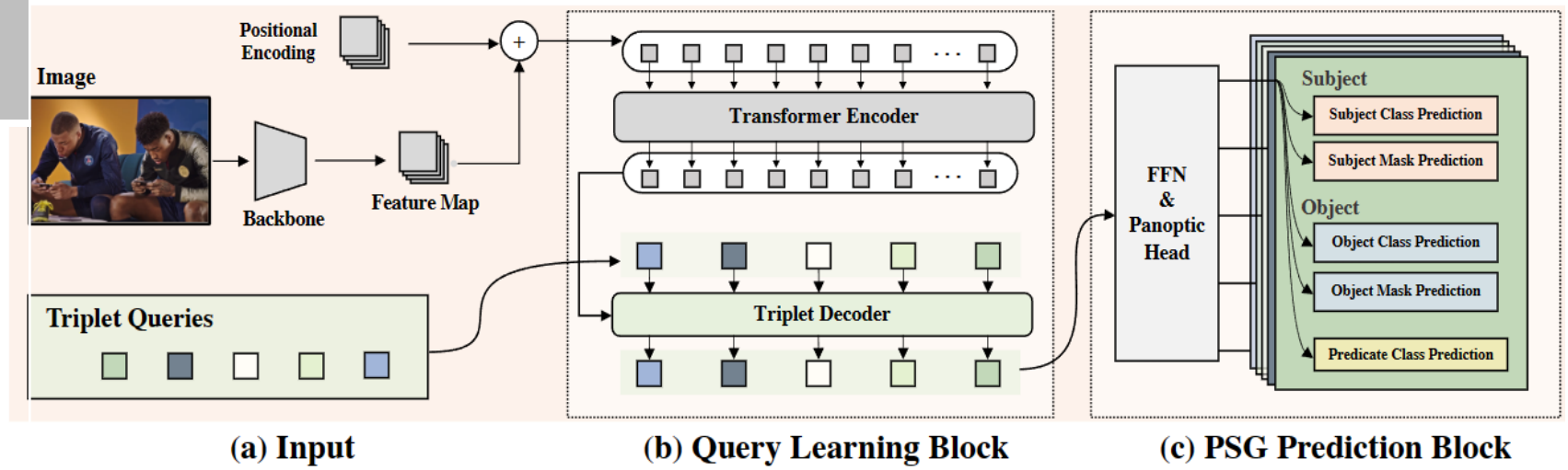
1. Object masks and Classes
2. Scene Graph



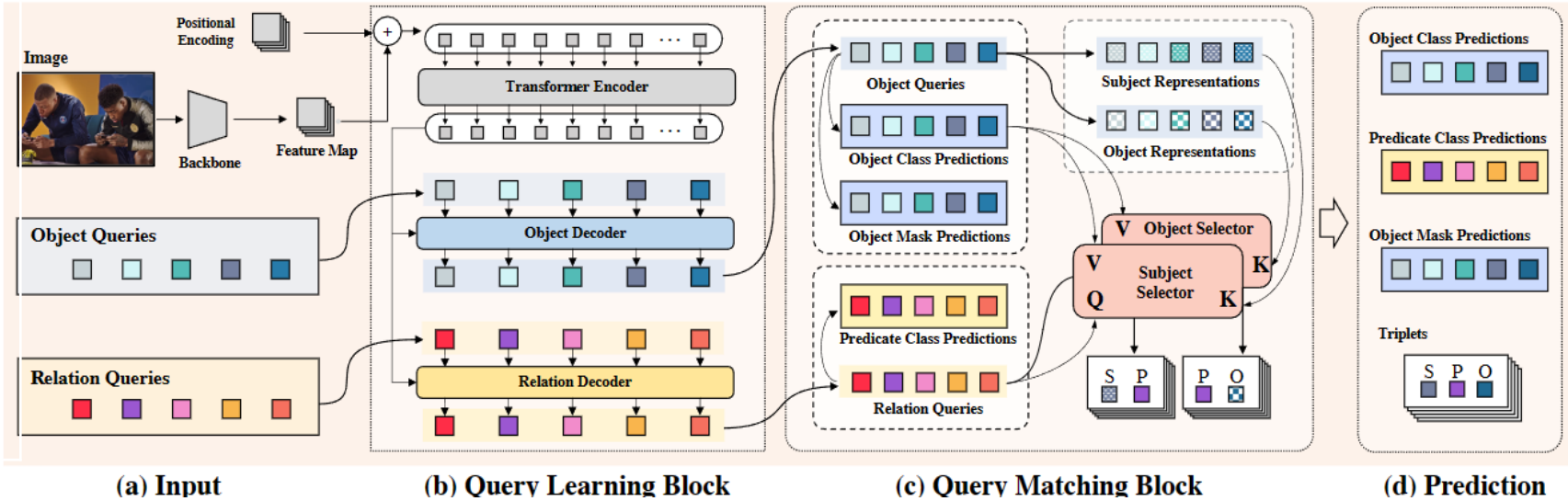


Previous Works

Method	drawback
PSGTR [1]	A single type query introduces instability in the learning process.

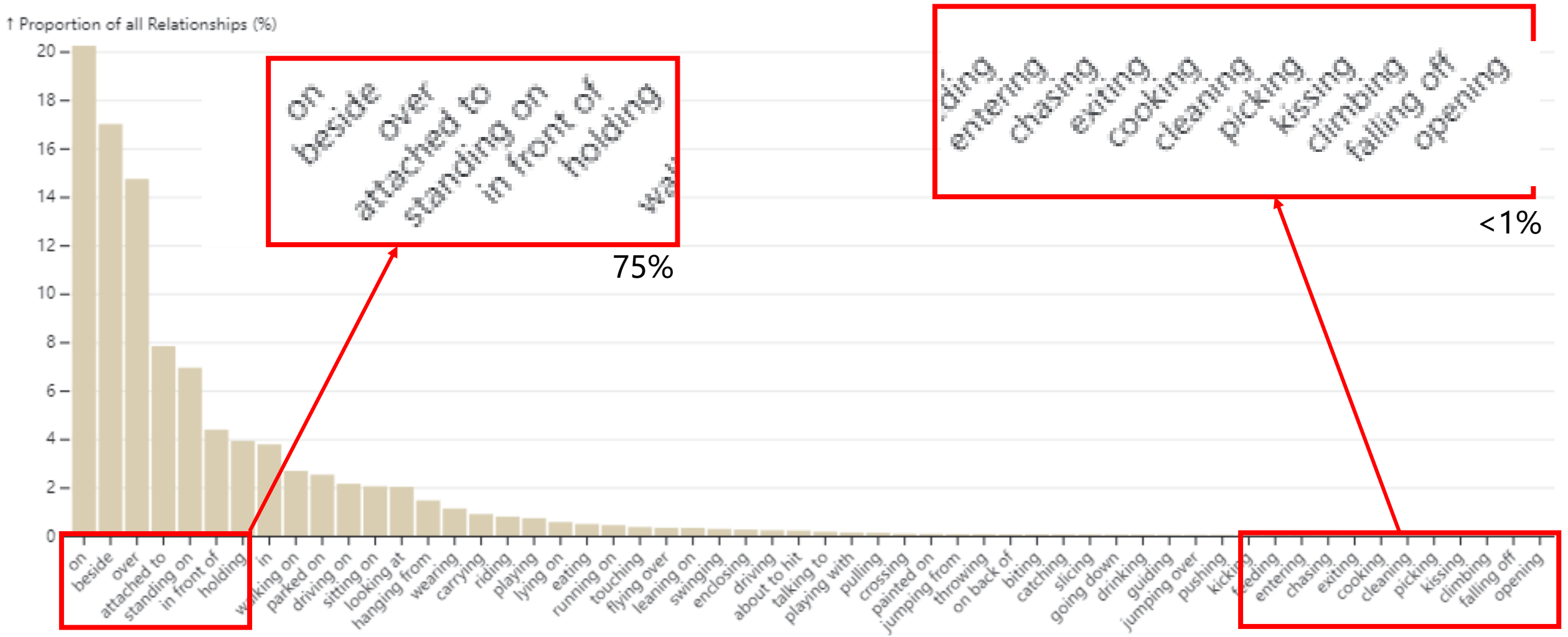


PSGFormer[1]	Easy matching make hindering the identification of the best match



[1] Jingkang Yang, Yi Zhe Ang, et al. **Panoptic scene graph generation**. In ECCV 2022

Long-tail problem



Relate prediction frequency

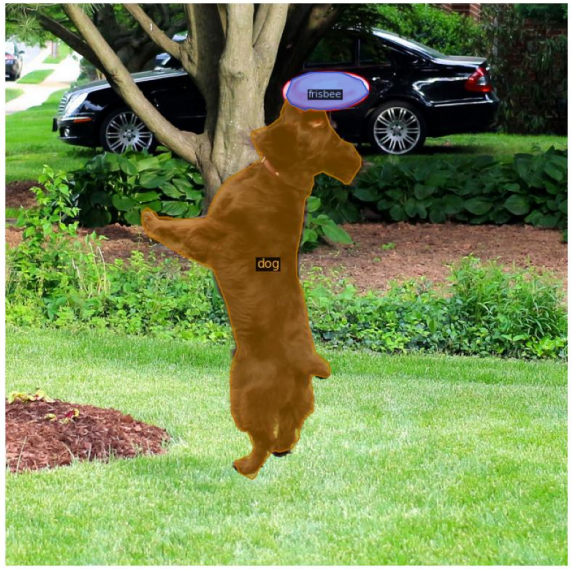
Long-tail problem

```
('opening', 9),
('falling off', 10),
('climbing', 14),
('kissing', 18),
('picking', 18),
('cleaning', 24),
('cooking', 24),
('exiting', 31),
('chasing', 52),
('entering', 66),
('feeding', 75),
('kicking', 77),
('pushing', 85),
('jumping over', 89),
('guiding', 96),
```

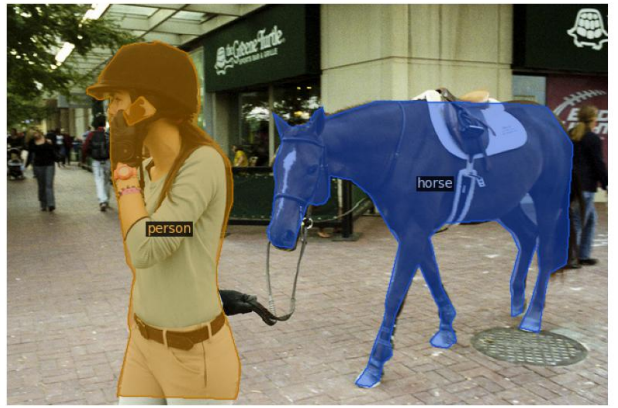
The 15 least frequent predicates

predicate	Rec100	predicate	Rec100	predicate	Rec100
over	0.5843	in front of	0.3074	beside	0.3648
on	0.3909	in	0.3053	attached to	0.2320
hanging from	0.3109	on back of	0.0256	falling off	0.0000
going down	0.0667	painted on	0.0000	walking on	0.5045
running on	0.5037	crossing	0.0000	standing on	0.5875
lying on	0.3146	sitting on	0.3743	flying over	0.2724
jumping over	0.0625	jumping from	0.0370	wearing	0.4314
holding	0.5658	carrying	0.3439	looking at	0.2630
guiding	0.0000	kissing	0.0000	eating	0.4137
drinking	0.2857	feeding	0.1111	biting	0.1667
catching	0.1333	picking	0.0000	playing with	0.0938
chasing	0.0000	climbing	0.0000	cleaning	0.0000
playing	0.6245	touching	0.1078	pushing	0.0000
pulling	0.0455	opening	0.0000	cooking	0.0000
talking to	0.0000	throwing	0.3333	slicing	0.2083
driving	0.2791	riding	0.5246	parked on	0.4815
driving on	0.6174	about to hit	0.5556	kicking	0.2500
swinging	0.7368	entering	0.0000	exiting	0.0000
enclosing	0.0902	leaning on	0.0327	None	None

12 of them have zero accuracy



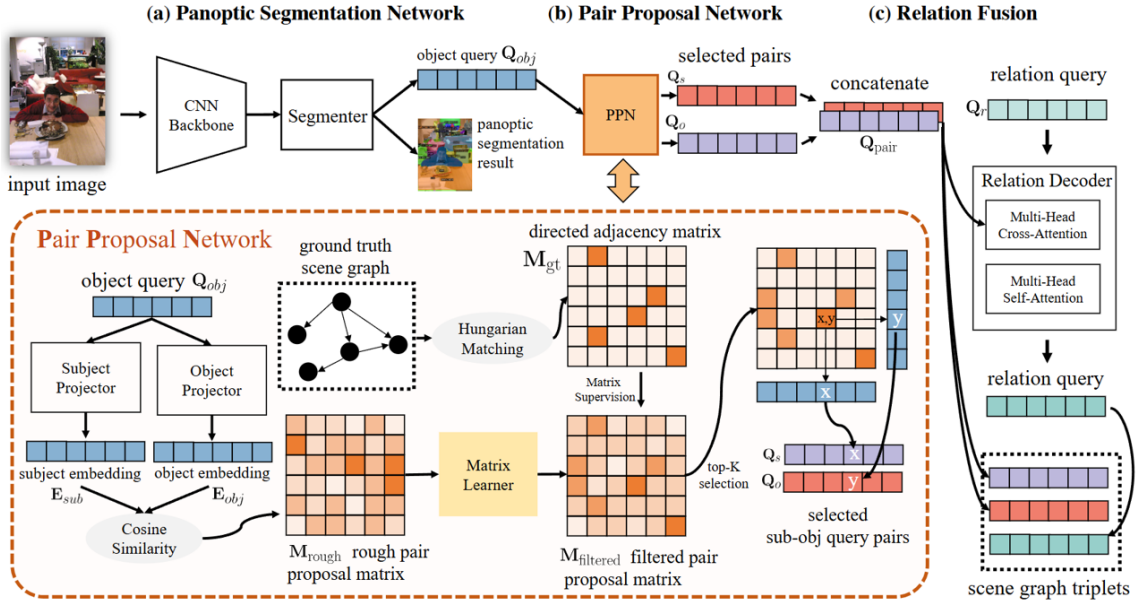
chasing

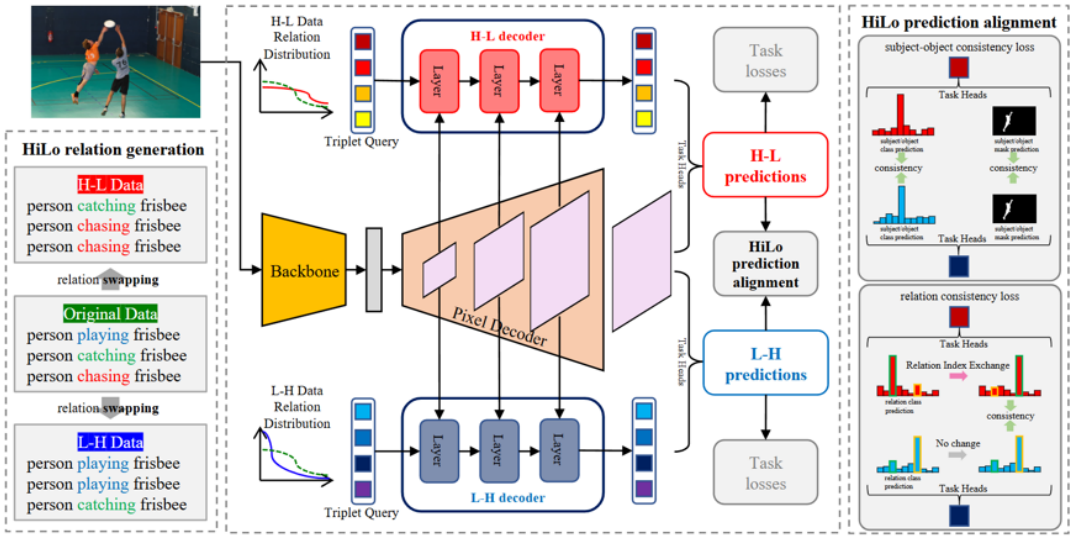


guiding

Previous Works

Method	drawback
PairNet[4]	The accuracy of the model is low, and the long tail problem is solved by weight adjustment loss.
Hilo[5]	Use additional relational data.



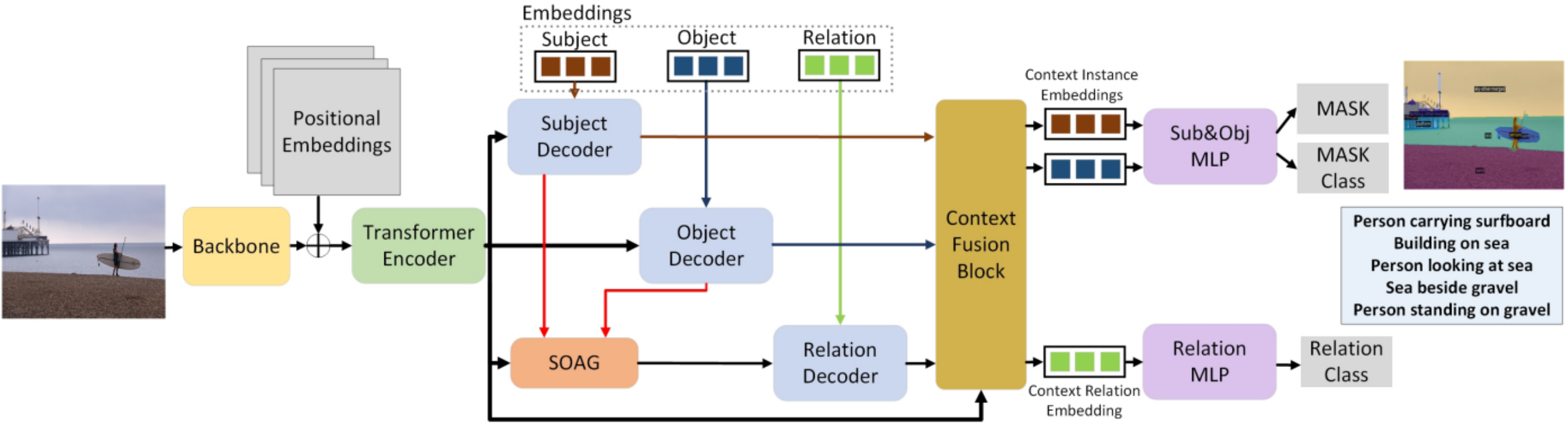


[4] Zijian Zhou, et al. Hilo: Exploiting high low frequency relations for unbiased panoptic scene graph generation. ICCV, 2023
[5] Jinghao Wang , et al. Pair then relation: Pair-net for panoptic scene graph generation.arXiv preprint arXiv:2307.08699, 2023

Previous Works

➢ Contextual Associated Triplet Queries (CATQ) [2]

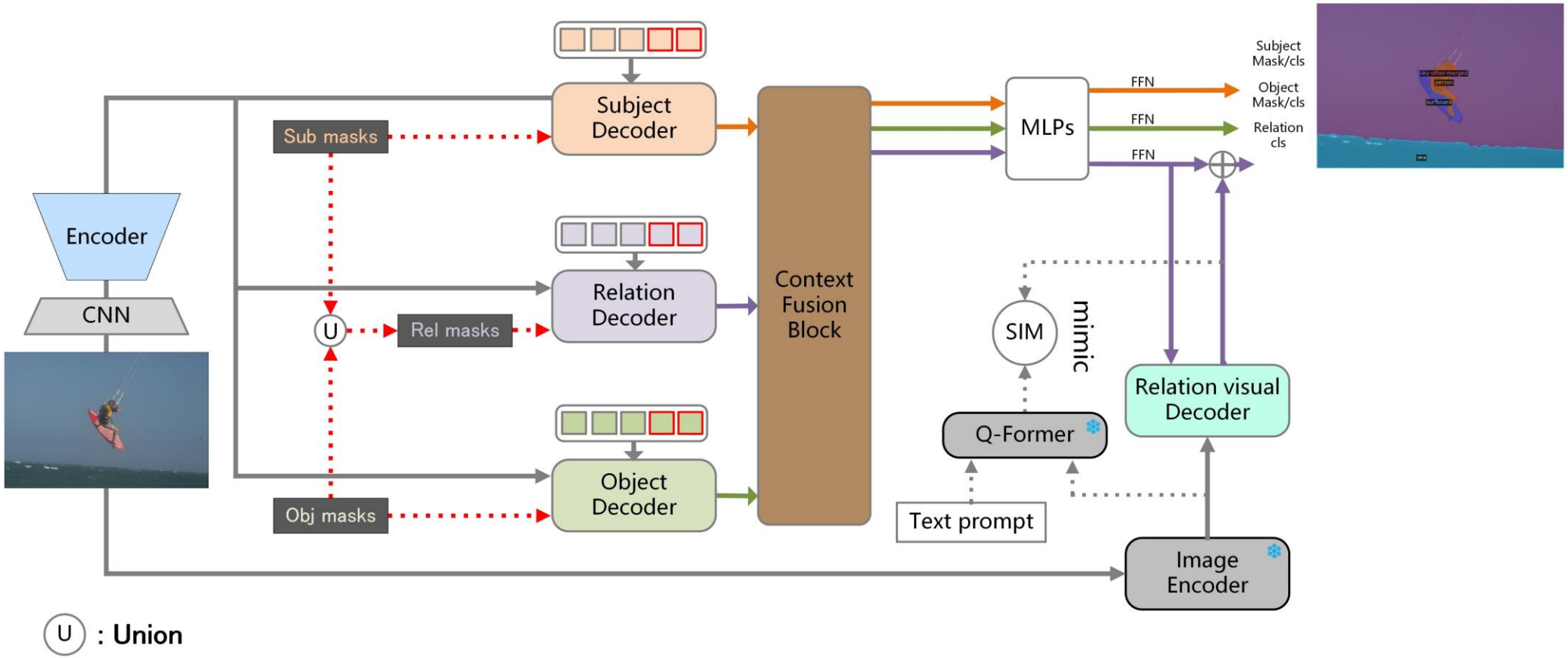
- **Triple decoding**
- ~~**Subject-Object Attention Guide (SOAG)**~~
- **Context Fusion Block (CFB)**



[2] Jingbin Xu, Junwen Chen, and Keiji Yanai. "**Contextual Associated Triplet Queries for Panoptic Scene Graph Generation**." *ACM Multimedia Asia* 2023.

Contributions

- **Contextual Associated Triplet Queries Relation Boosting (CATQ-RB)**
  - **Relation Boosting Module (RBM)**
  - **Relation Boosting Training (RBT)**



Relation Boosting Module

- Visual knowledge prompt
- Text guide



CATQ

Relation query

FFN

From CFB

Training

triplet context

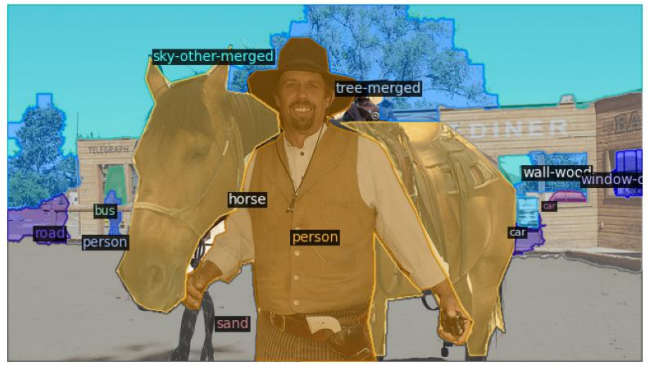
Comparative learning

Language features

Visual feature

Boosting

Visual-Language model



Person holding horse

Ground truth triplet:

- person guiding horse
- car parked on road
- bus parked on road
- horse walking on sand
- wall-wood in front of car
- sky over tree

Relation Boosting Module

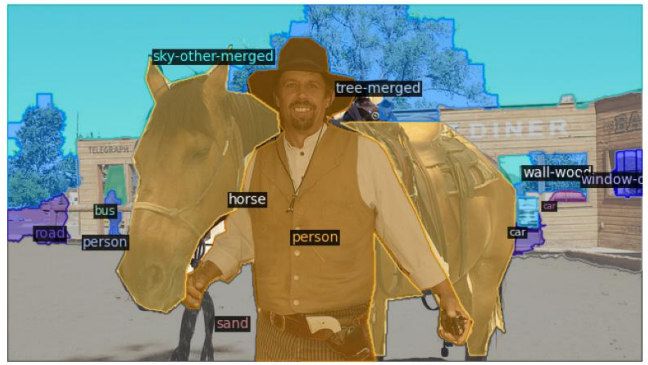
- Visual knowledge prompt



CATQ

Relation query

FFN



Person holding horse

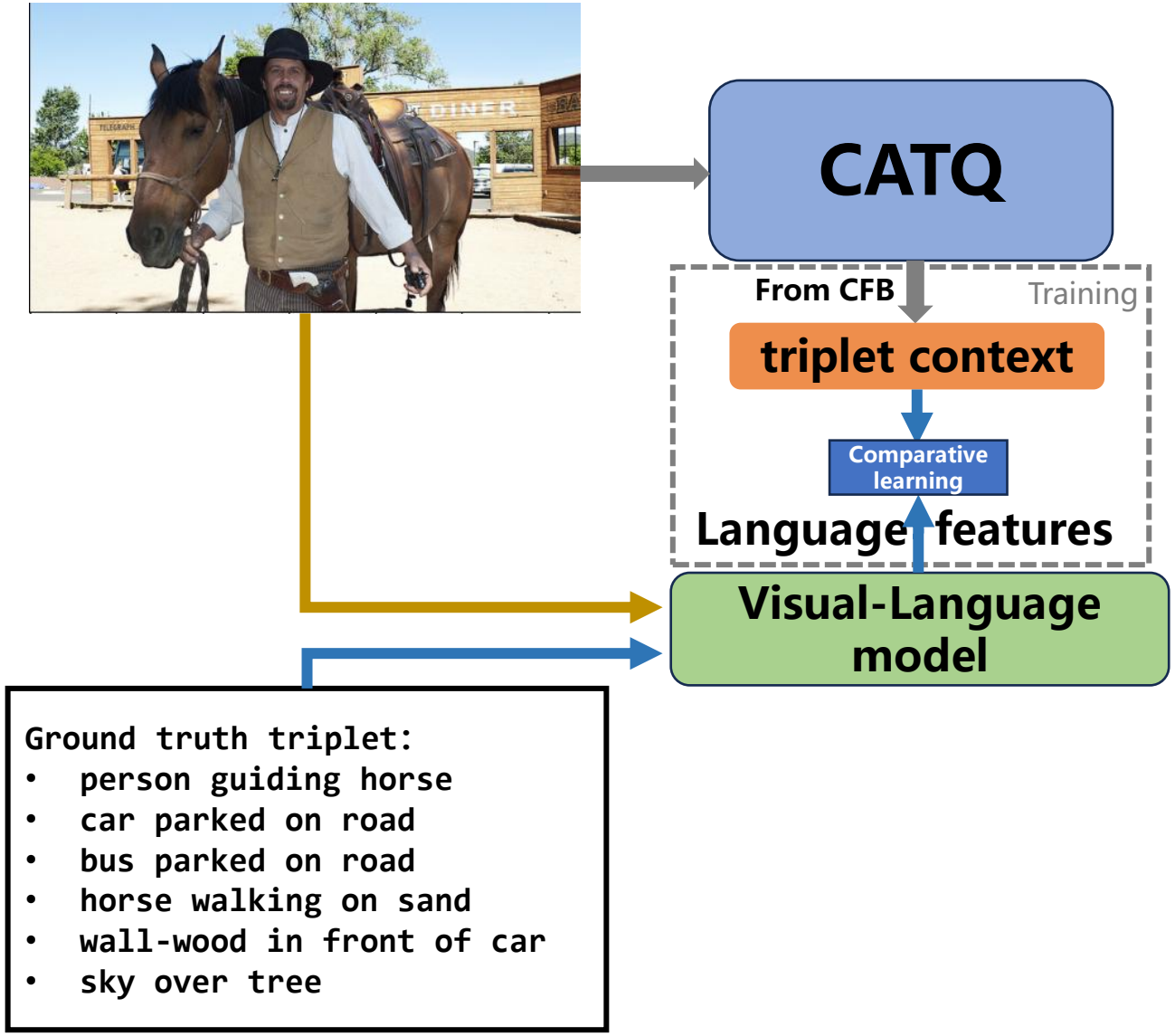
Boosting

Visual feature

Visual-Language model

Relation Boosting Module

- Text guide

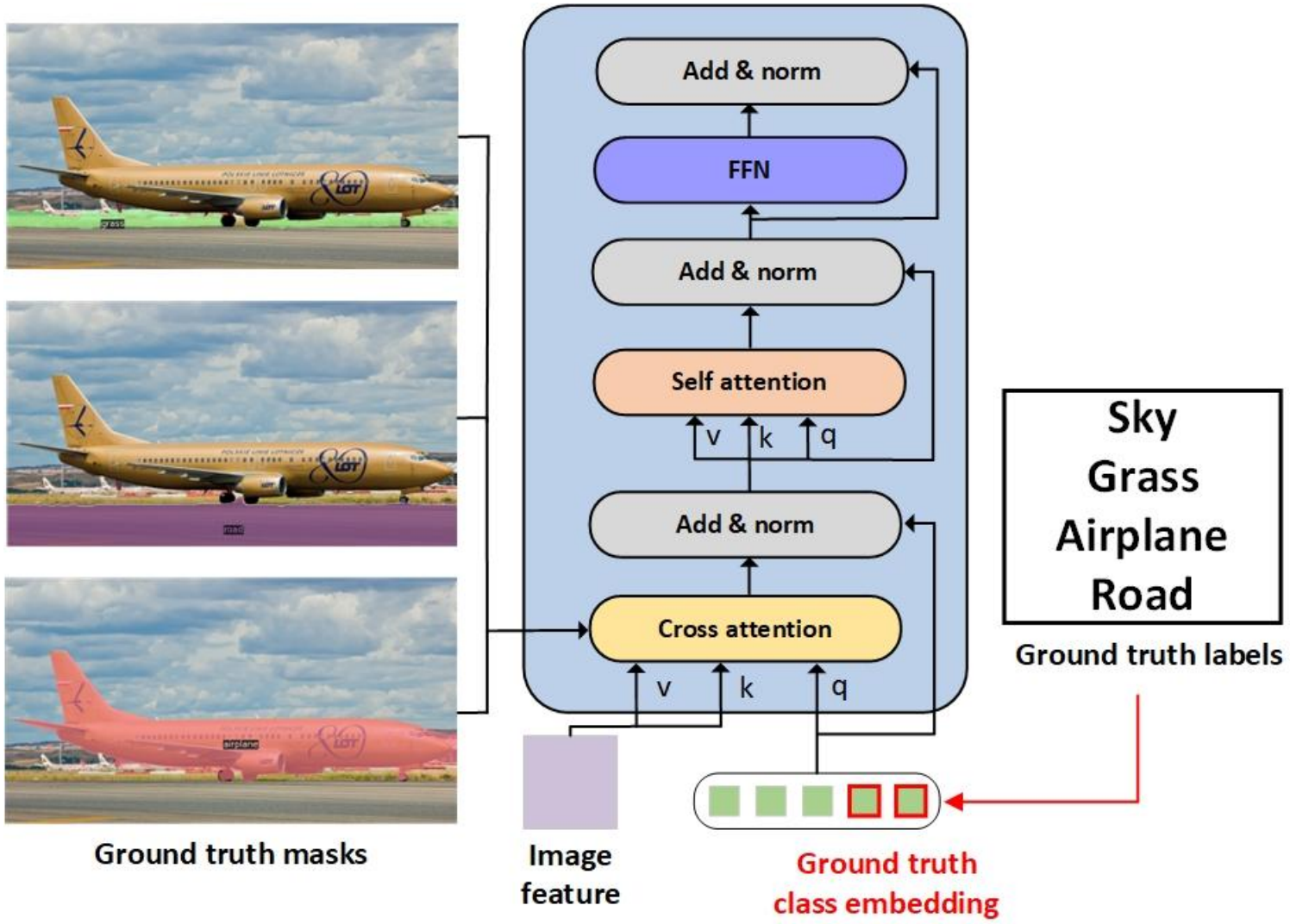


Relation Boosting Training



Relation Boosting Training

- Mask-lead training
- Label-lead training



Result

- **Comparison Between the Proposed Method and the PSG Baseline Method.**

Method	Backbone	*R@20*	*R@50*	*R@100*	*mR@20*	*mR@50*	*mR@100*	*PQ*
Two-stage								
IMP [17]	ResNet-50	16.5	18.2	18.6	6.5	7.1	7.2	40.2
MOTIFS [21]	ResNet-50	20.0	21.7	22.0	9.1	9.6	9.7	40.2
VCTree [14]	ResNet-50	20.6	22.1	22.5	9.7	10.2	10.2	40.2
GPSNet [10]	ResNet-50	17.8	19.6	20.1	7.0	7.5	7.7	40.2
One-stage								
PSGTR [19]	ResNet-50	28.4	34.4	36.3	16.6	20.8	22.1	13.9
PSGFormer [19]	ResNet-50	18.0	19.6	20.1	14.8	17.0	17.6	36.8
PairNet [16]	ResNet-50	29.6	35.6	39.6	**24.7**	28.5	30.6	-
HiLo [23]	ResNet-50	34.1	**40.7**	**43.0**	23.7	30.3	**33.1**	41.6
CATQ [18]	ResNet-50	**34.8**	39.7	40.3	20.9	24.9	25.2	35.9
CATQ-RB(ours)	ResNet-50	**34.8**	40.4	41.2	23.8	**30.8**	32.1	35.9

2.9 ↑ 5.9 ↑ 6.9 ↑

Results

- **Comparison Between the Proposed Method and the PSG Baseline Method.**

Method	Backbone	*R@20*	*R@50*	*R@100*	*mR@20*	*mR@50*	*mR@100*	*PQ*
Two-stage								
IMP [17]	ResNet-50	16.5	18.2	18.6	6.5	7.1	7.2	40.2
MOTIFS [21]	ResNet-50	20.0	21.7	22.0	9.1	9.6	9.7	40.2
VCTree [14]	ResNet-50	20.6	22.1	22.5	9.7	10.2	10.2	40.2
GPSNet [10]	ResNet-50	17.8	19.6	20.1	7.0	7.5	7.7	40.2
One-stage								
PSGTR [19]	ResNet-50	28.4	34.4	36.3	16.6	20.8	22.1	13.9
PSGFormer [19]	ResNet-50	18.0	19.6	20.1	14.8	17.0	17.6	36.8
PairNet [16]	ResNet-50	29.6	35.6	39.6	**24.7**	28.5	30.6	-
HiLo [23]	ResNet-50	34.1	**40.7**	**43.0**	23.7	30.3	**33.1**	41.6
CATQ [18]	ResNet-50	**34.8**	39.7	40.3	20.9	24.9	25.2	35.9
CATQ-RB(ours)	ResNet-50	**34.8**	40.4	41.2	23.8	**30.8**	32.1	35.9

0.1 ↑ 0.5 ↑

Experiments

- **Application of Weight-Adjusted Loss**

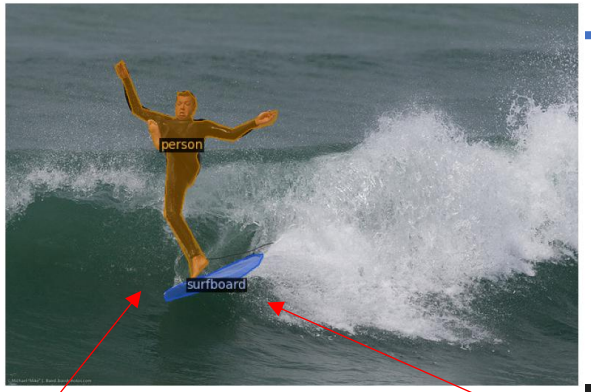
Model	mRecall@20	mRecall@50	mRecall@100
CATQ-RB	20.0	**33.2**	**38.8**
PairNet	**24.7**	28.5	30.6
HiLo	23.7	30.3	33.1

- **Text guide**

Ablation	mRecall@20	mRecall@50	mRecall@100
w/ Visual Information	**22.5**	**30.0**	**31.0**
w/o Visual Information	20.4	24.1	24.4
w/o text	21.7	26.8	27.0

Experiments

➢ Comparison of Predicate Accuracy



predicate	Rec100	predicate	Rec100	predicate	Rec100
over	0.5843	in front of	0.3074	beside	0.3648
on	0.3909	in	0.3053	attached to	0.2320
hanging from	0.3109	on back of	0.0256	falling off	0.0000
going down	0.0667	painted on	0.0000	walking on	0.5045
running on	0.5037	crossing	0.0000	standing on	0.5875
lying on	0.3146	sitting on	0.3743	flying over	0.2724
jumping over	0.0625	jumping from	0.0370	wearing	0.4314
holding	0.5658	carrying	0.3439	looking at	0.2630
guiding	0.0000	kissing	0.0000	eating	0.4137
drinking	0.2857	feeding	0.1111	biting	0.1667
catching	0.1333	picking	0.0000	playing with	0.0938
chasing	0.0000	climbing	0.0000	cleaning	0.0000
playing	0.6245	touching	0.1078	pushing	0.0000
pulling	0.0455	opening	0.0000	cooking	0.0000
talking to	0.0000	throwing	0.3333	slicing	0.2083
driving	0.2791	riding	0.5246	parked on	0.4815
driving on	0.6174	about to hit	0.5556	kicking	0.2500
swinging	0.7368	entering	0.0000	exiting	0.0000
enclosing	0.0902	leaning on	0.0327	None	None

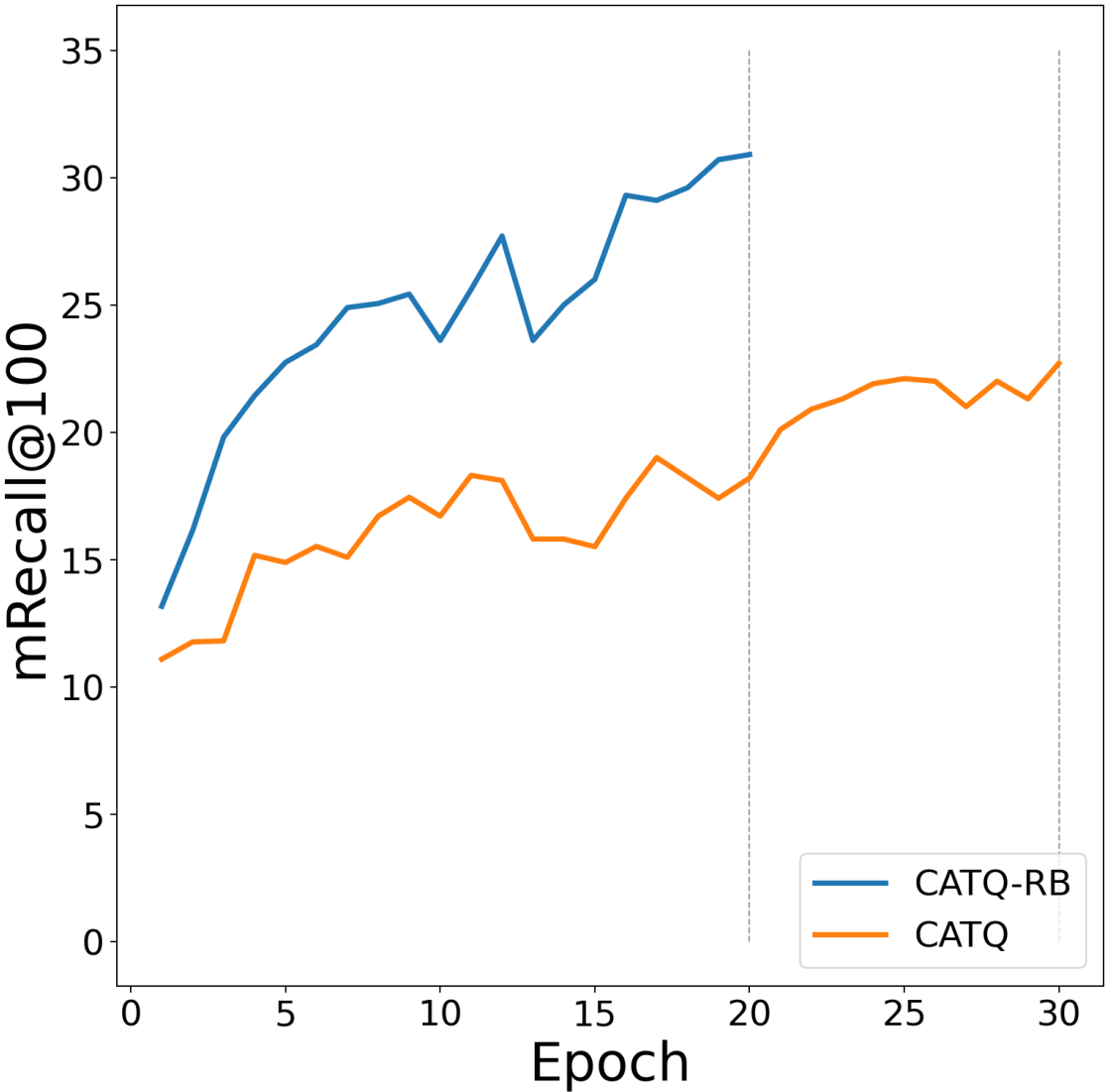
predicate	Rec100	predicate	Rec100	predicate	Rec100
over	0.5553	in front of	0.3574	beside	0.2864
on	0.3744	in	0.3678	attached to	0.2686
hanging from	0.2945	on back of	0.0641	falling off	0.3333
going down	0.2667	painted on	0.0000	walking on	0.5000
running on	0.6013	crossing	0.2143	standing on	0.5641
lying on	0.3545	sitting on	0.4230	flying over	0.4064
jumping over	0.3125	jumping from	0.5000	wearing	0.4451
holding	0.5846	carrying	0.4749	looking at	0.3558
guiding	0.1538	kissing	1.0000	eating	0.5234
drinking	0.4286	feeding	0.4444	biting	0.5000
catching	0.4000	picking	0.0000	playing with	0.2969
chasing	0.2500	climbing	0.2500	cleaning	0.1429
playing	0.5963	touching	0.3235	pushing	0.2917
pulling	0.3182	opening	0.0000	cooking	0.4000
talking to	0.1579	throwing	0.7500	slicing	0.5833
driving	0.4302	riding	0.6128	parked on	0.4910
driving on	0.6204	about to hit	0.8148	kicking	0.7500
swinging	0.8158	entering	0.2273	exiting	0.0000
enclosing	0.2295	leaning on	0.0000	None	None

CATQ
12 are zero

CATQ-RB
3 are zero

Experiments

➢Training Convergence Time





Image

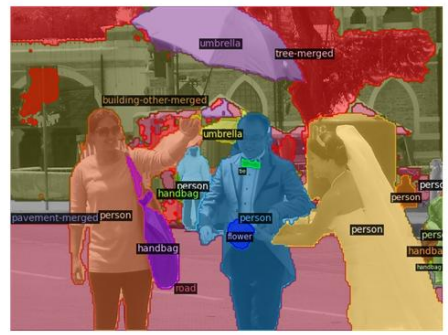
person sitting on road
person looking at person
person looking at person
person holding umbrella
person carrying handbag
person wearing tie
person looking at flower
person picking flower
person walking on road
person carrying handbag
person walking on pavement
person sitting on pavement
person holding flower
banner hanging from building
tree in front of building

Ground Turth



Person wearing tie
Pavement attached to road
Person holding umbrella
Person walking on road
Umbrella over person
Person beside person
Person standing on road
Person walking on road
Person carrying handbag
Road attached to pavement
Person holding umbrella
Person walking on road
Umbrella over person
Person standing on road
Person holding flower
Person walking on road
Person standing on road
Person holding umbrella
Person wearing tie
Person holding umbrella

Prediction

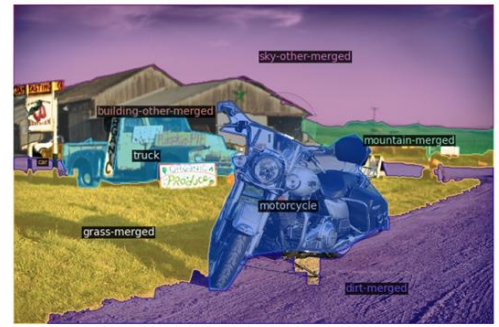




Image

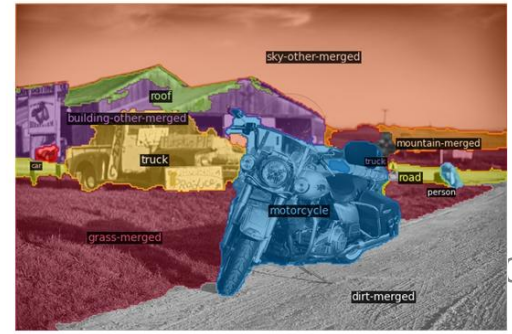
motorcycle in front of truck
motorcycle beside grass
motorcycle parked on dirt
truck in front of building
sky over dirt
dirt in front of truck
dirt beside truck

Ground Turth



Glass attached to dirt
Building on grass
Person on dirt
Roof on building
Building beside truck
Sky over building
Truck beside building
Truck in front of building
Truck parked on dirt
Motorcycle beside grass
Sky over person
Sky over roof
Sky over mountain
Motorcycle on dirt
Sky over truck
Sky over dirt
Truck parked on grass
Sky over motorcycle
Sky over grass
Motorcycle parked on dirt

Prediction



Conclusion

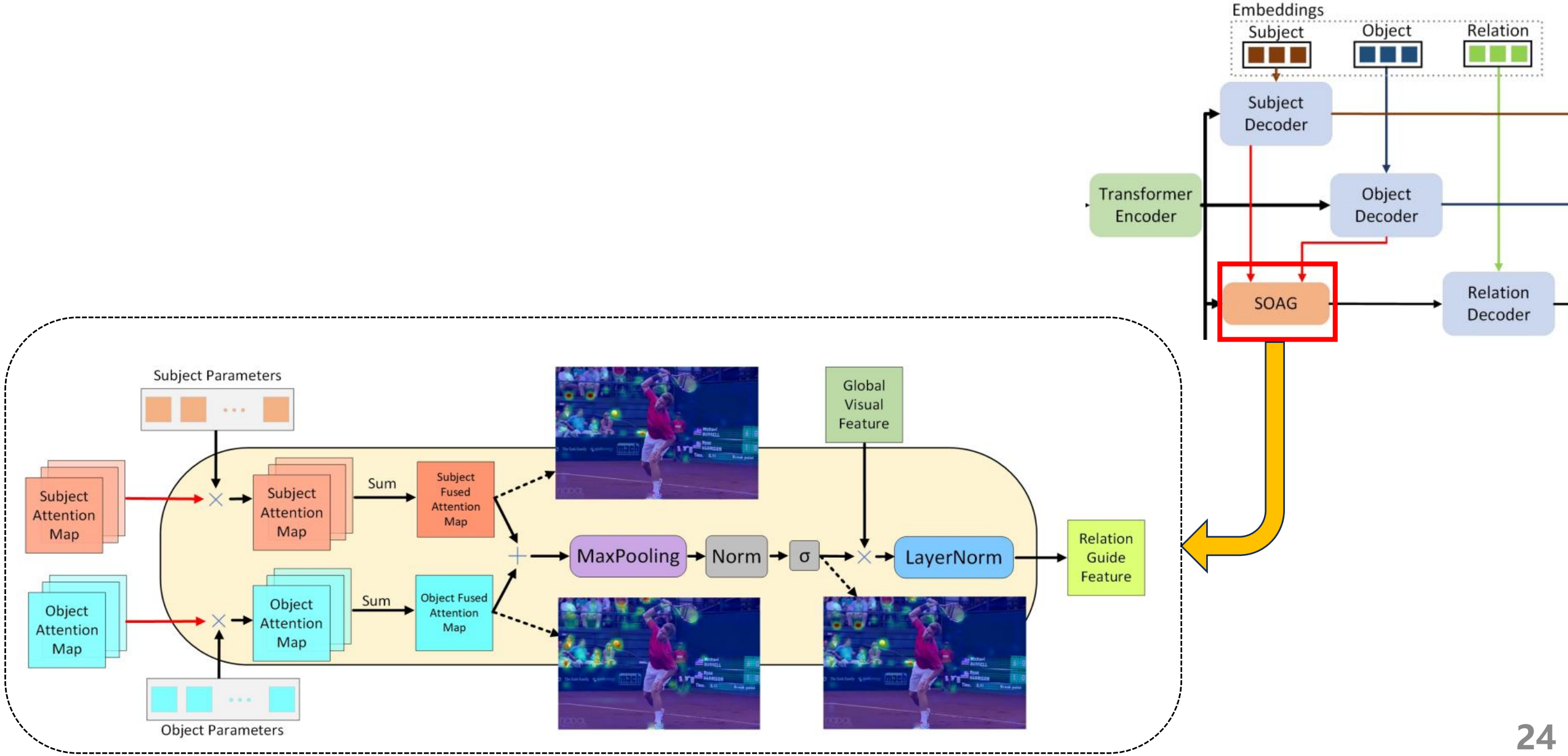
- **We propose CATQ-RB to address the long-tail problem by leveraging large-scale vision-language models and introduce the RBT strategy to mitigate instability in PSG training.**
  1. **RBM:**
     - Visual knowledge prompt
     - Text guide
  2. **RBT:**
     - Mask-lead training
     - Label-lead training
- **Experiments confirm that our model effectively addresses the long-tail problem while improving training efficiency.**

Thanks for your listening



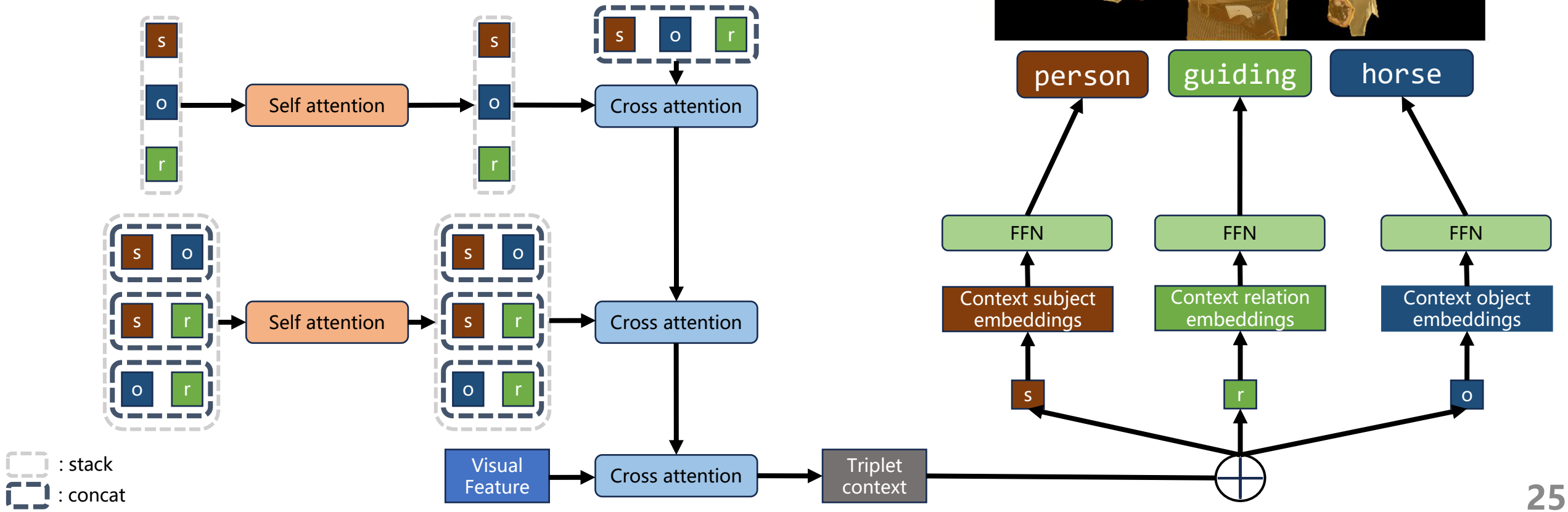
Supplementary

Subject-Object Attention Guide (SOAG)



Context Fusion Block (CFB)



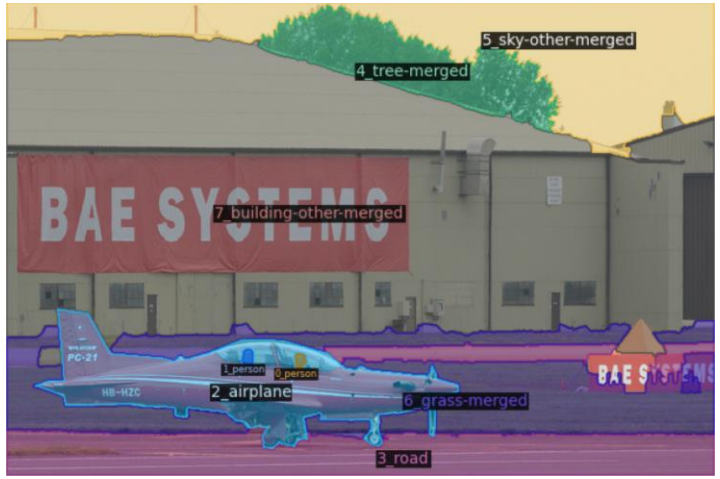


OpenPSG Dataset

- 48749 images
- 56 relation classes
- 133 instance classes(80 object、53 stuff)



image



Segment Annotations

0 person **driving** 2 airplane

1 person **in** 2 airplane

2 airplane **on** 3 road

2 airplane **on** 3 road

5 sky-other-merged **over** 7 building-other-merged

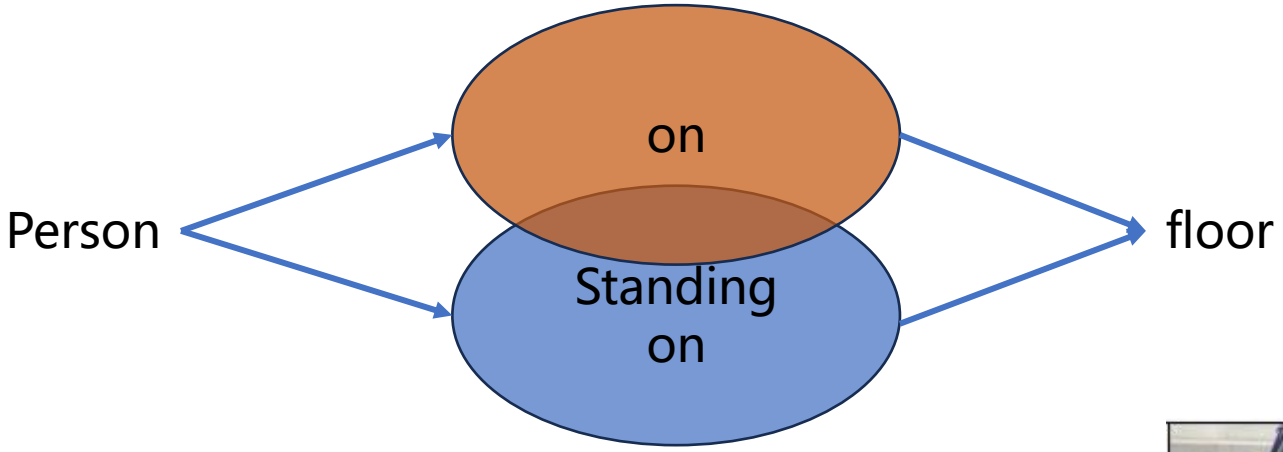
7 building-other-merged **in front of** 4 tree-merged

7 building-other-merged **in front of** 4 tree-merged

Scene graph Annotations

Limitations of Using Only Linguistic Knowledge

➢SGG task long tail issue —reason→ Overlapping Semantics



➢Diversity within Each Relationship Category

- Person looking at frisbee
- Person throwing frisbee



Instability in Query Prediction


(a)


(b)

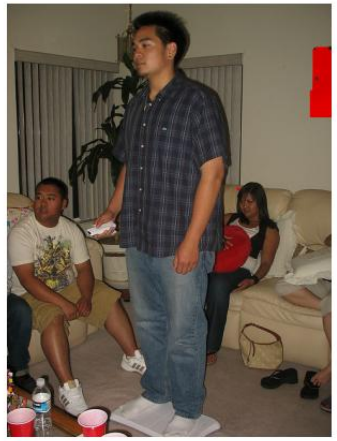

(c)


(d)


(e)


(f)


(g)

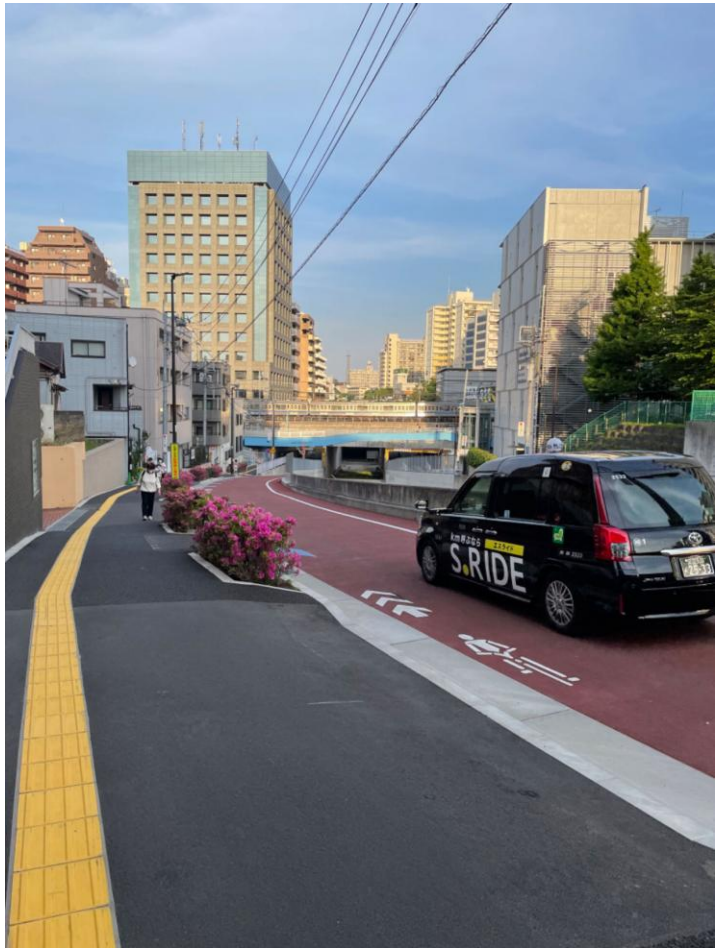

(h)

The figure is cited from [6]

[6] Zhang, Hao, et al. "MP-Former: Mask-piloted transformer for image segmentation." CVPR 2023.

Experiment

- **Result Visualization (Real Images)**



Original image



Segment result

Sky over bridge
Sky over building
Sky over potted plant
Road attached to pavement
Sky over fence
Sky over wall
Car on road
Sky over car
Sky over person
Person beside potted plant
Sky over potted plant
Person beside person
Car driving on road
Sky over fence
Person standing on pavement
Person on pavement
Sky over building
Person walking on pavement
Sky over pavement
Potted plant on pavement