Think before Placement: Common Sense Enhanced Transformer for Object Placement

Yaxuan Qin, Jiayu Xu, Ruiping Wang, Xilin Chen

Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
{qinyaxuan21s, xujiayu22s, wangruiping, xlchen}@ict.ac.cn

code and dataset available

MOTIVATION

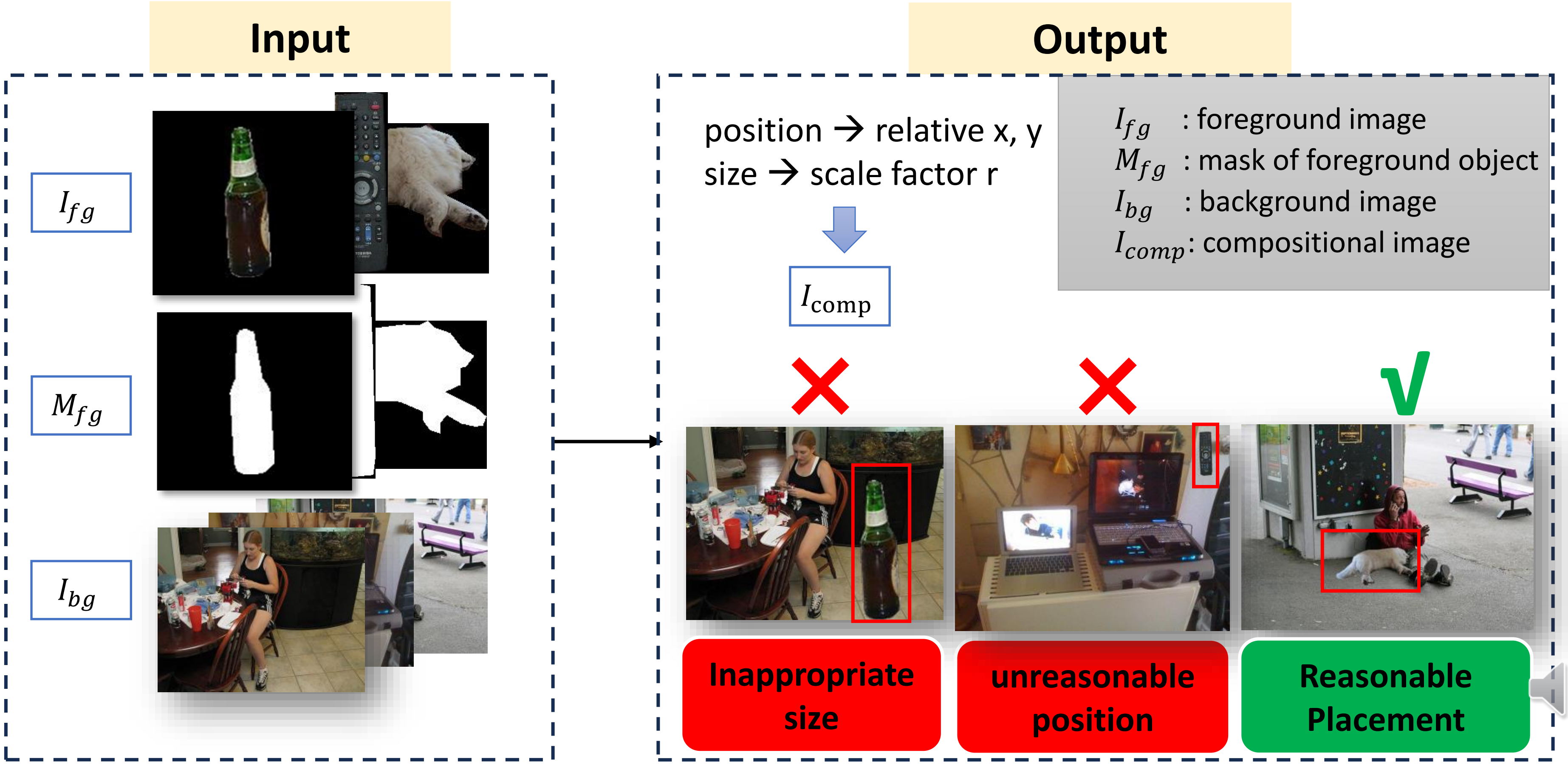
➢ **Image Composition**: paste a **foreground object** from one image on another **background image**

- ❑ **Object placement:** size & position
- ❑ **Image blending:** natural boundary
- ❑ **Image harmonization:** illumination statistics

art · entertainment · data augmentation

MOTIVATION

➢Object Placement



Input

I_{fg}

M_{fg}

I_{bg}

Output

position → relative x, y
size → scale factor r

I_{comp}

I_{fg} : foreground image
M_{fg} : mask of foreground object
I_{bg} : background image
I_{comp}: compositional image

✗ Inappropriate size

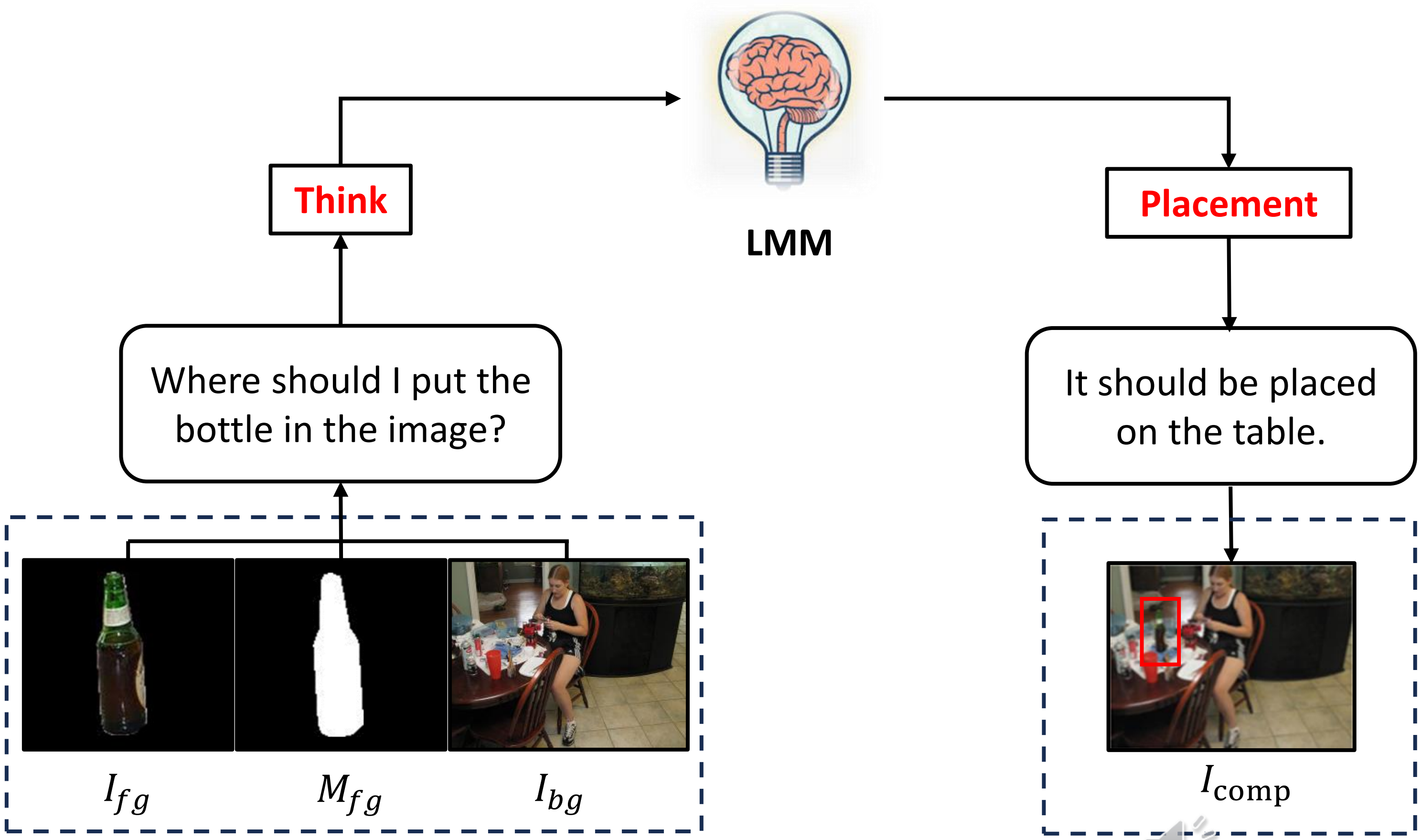
✗ unreasonable position

√ Reasonable Placement

MOTIVATION

- **Think before Placement**
  - ❑ **Think**： generate the guiding caption
  - ❑ **Placement**： predict the suitable position and size based on the result of “Think” process

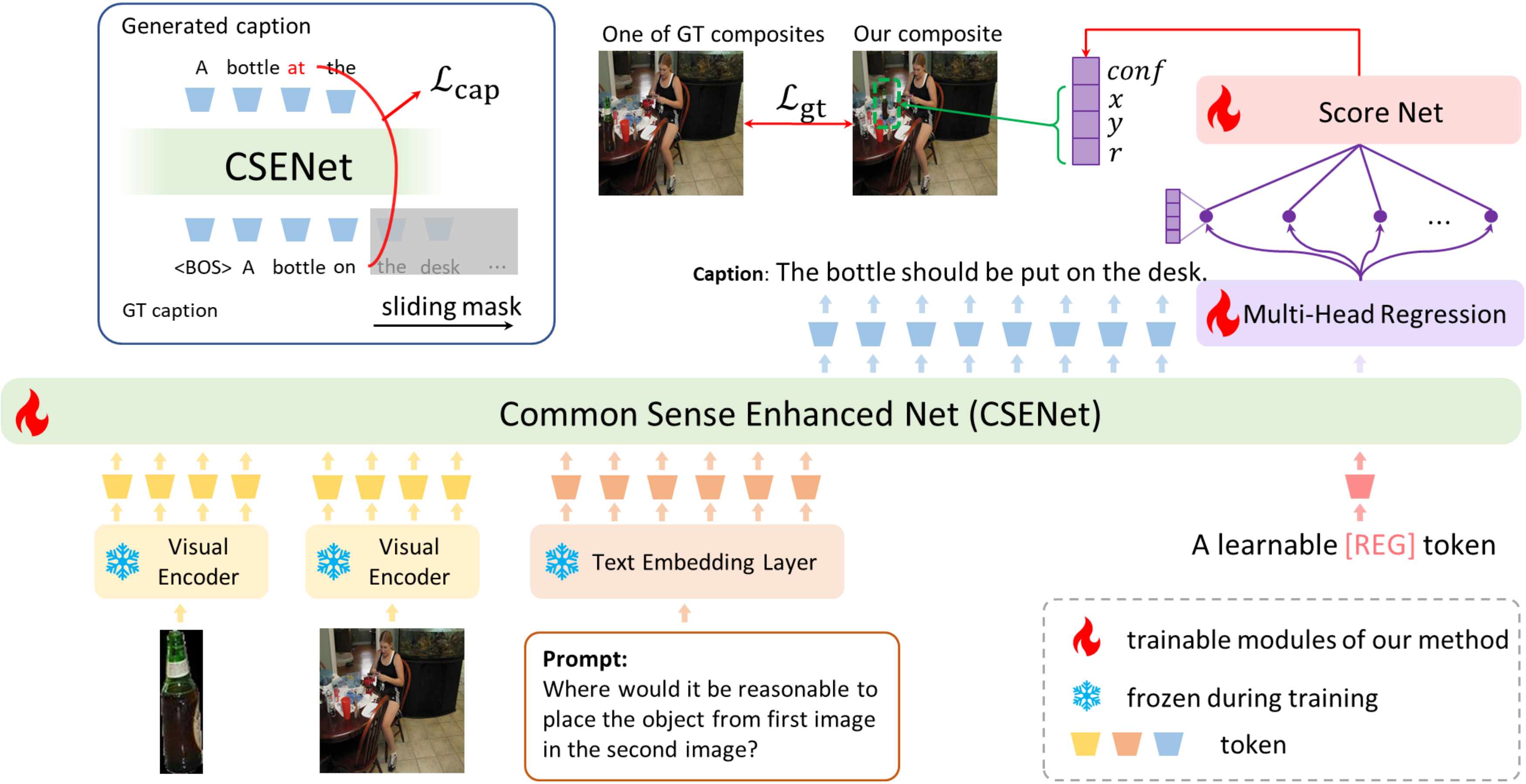
Common sense enhanced



Think
LMM
Placement
Where should I put the bottle in the image?
It should be placed on the table.
I_{fg}
M_{fg}
I_{bg}
I_{comp}

FRAMEWORK

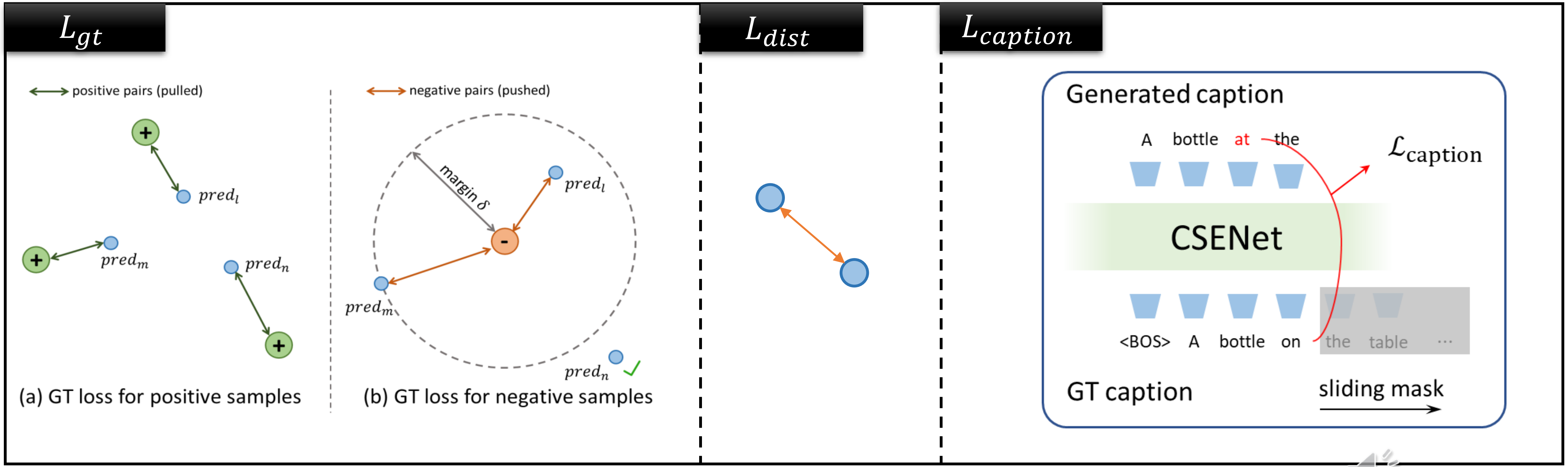
➢LLM decoder + multi-head regression + score net



Generated caption
A bottle at the
$\mathcal{L}_{\text{cap}}$
CSENet
<BOS> A bottle on the desk ···
GT caption
sliding mask
One of GT composites
Our composite
$\mathcal{L}_{\text{gt}}$
$conf$
x
y
r
Score Net
···
Caption: The bottle should be put on the desk.
Multi-Head Regression
Common Sense Enhanced Net (CSENet)
Visual Encoder
Visual Encoder
Text Embedding Layer
A learnable [REG] token
Prompt:
Where would it be reasonable to place the object from first image in the second image?
trainable modules of our method
frozen during training
token

LOSS DESIGN

- $L_{gt}+L_{dist}+L_{caption}$
  - ❑Ground truth Loss + Distance Loss + Caption Loss



(a) GT loss for positive samples

(b) GT loss for negative samples

OPAZ dataset

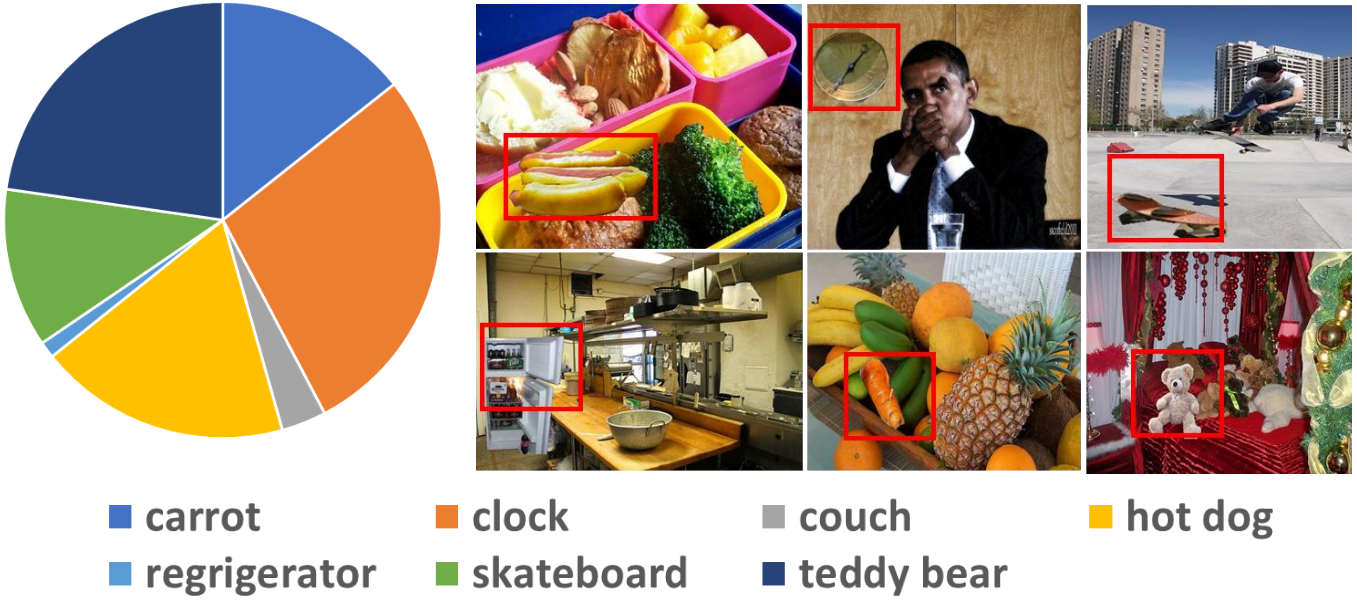
➢ **OPAZ dataset:** We construct an evaluation dataset to test the zero-shot transfer capabilities for the object placement task

❑ **8,160** generated image, of which **1,390** are rational and **6770** are irrational

❑ About 7 distinct categories of foreground objects

❑ About 15 representative images for each category

❑ About 10 suitable background images for each category



EVALUATION

➢Experiments on OPA dataset

❑Comparisons with baselines

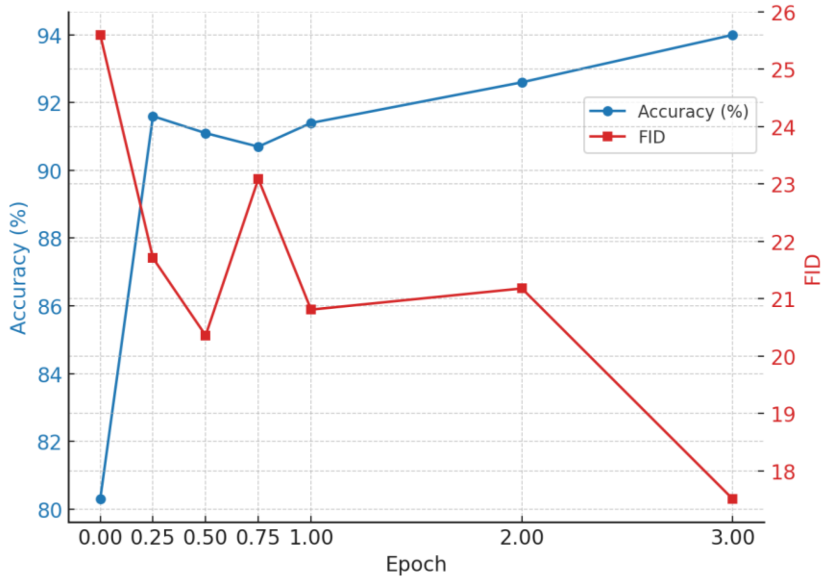
Model	User Study↑	Accuracy↑	FID↓	Mean IoU ↑	LPIPS ↑
TopNet [CVPR'23]	0.072	44.7	28.81	0.227	0.110
TERSE [CVPR'19]	0.096	67.9	46.94	0.171	0
PlaceNet [ECCV'20]	0.140	68.3	36.69	0.277	0.160
GracoNet [ECCV'22]	0.192	84.7	27.75	**0.336**	0.206
IOPRE [ICML'23]	0.234	89.5	21.59	0.226	**0.214**
CSENet (Ours)	**0.266**	**94.0**	**17.51**	0.321	0.137

EVALUATION

➢Experiments on OPA dataset

❑Ablations on the backbone and guiding captions

PT	Caption	#Heads	Accuracy↑	FID↓
×	w/o	1	72.9	49.45
√	w/o	1	80.9	39.89
√	w/o	10	90.8	22.72
√	Simple	10	91.3	20.07
√	Detailed	10	**94.0**	**17.51**

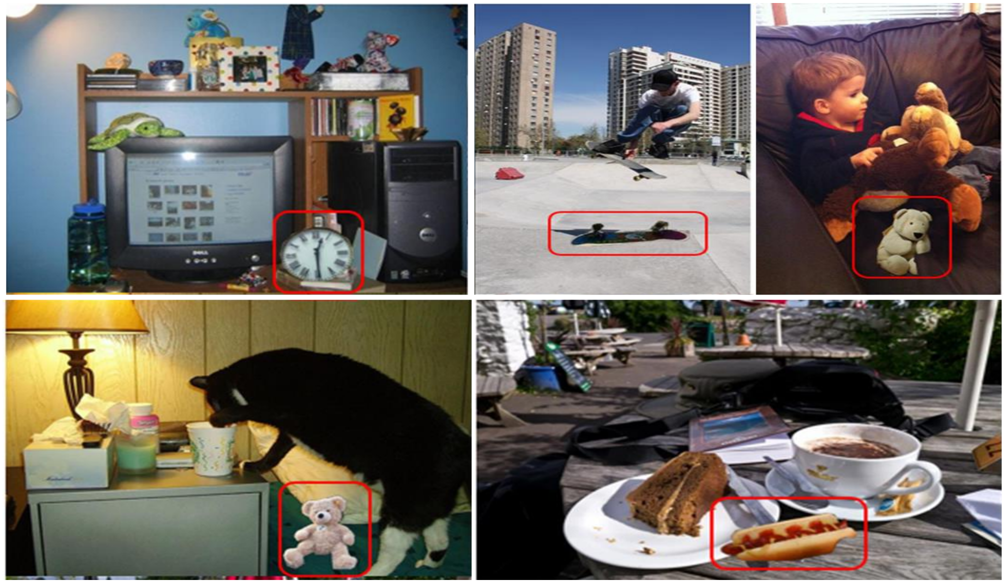


EVALUATION

➢Experiments on OPAZ dataset

❑zero-shot setting

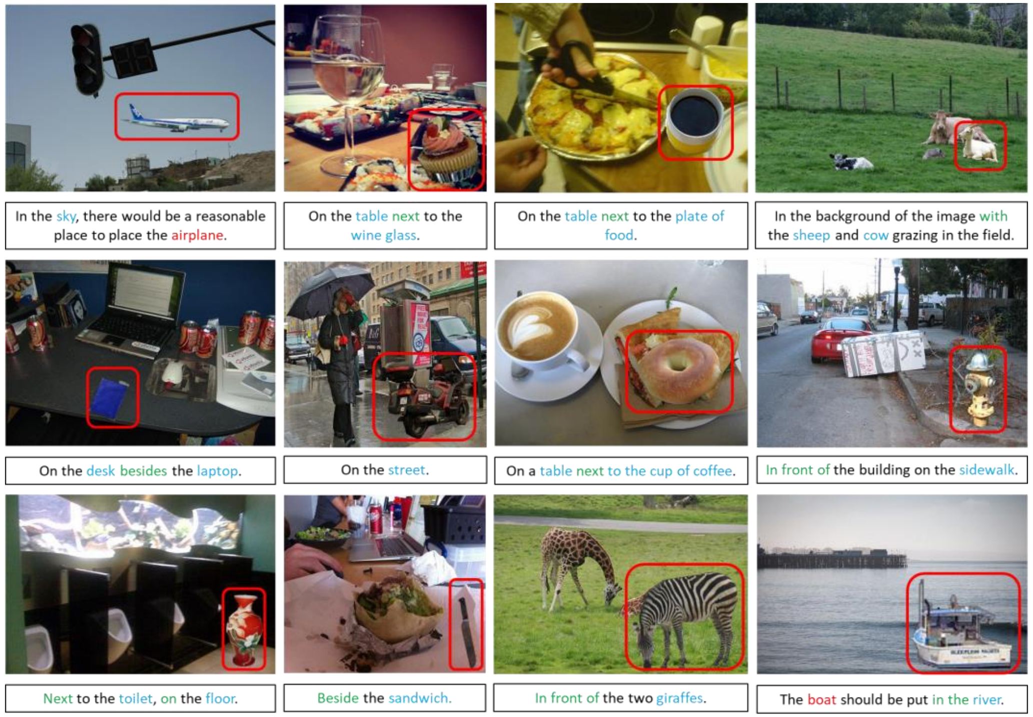
	User Study↑	Accuracy↑	FID↓
TopNet [CVPR'23]	0.096	18.9	67.7
TERSE [CVPR'19]	0.115	34.0	81.1
PlaceNet [ECCV'20]	0.142	36.7	63.6
GracoNet [ECCV'22]	0.174	43.1	59.1
IOPRE [ICML'23]	0.221	58.6	**27.8**
CSENet (Ours)	**0.251**	**61.8**	42.1



EVALUATION

➢Visualization

❑Comparisons with baselines



In the sky, there would be a reasonable place to place the airplane.

On the table next to the wine glass.

On the table next to the plate of food.

In the background of the image with the sheep and cow grazing in the field.

On the desk besides the laptop.

On the street.

On a table next to the cup of coffee.

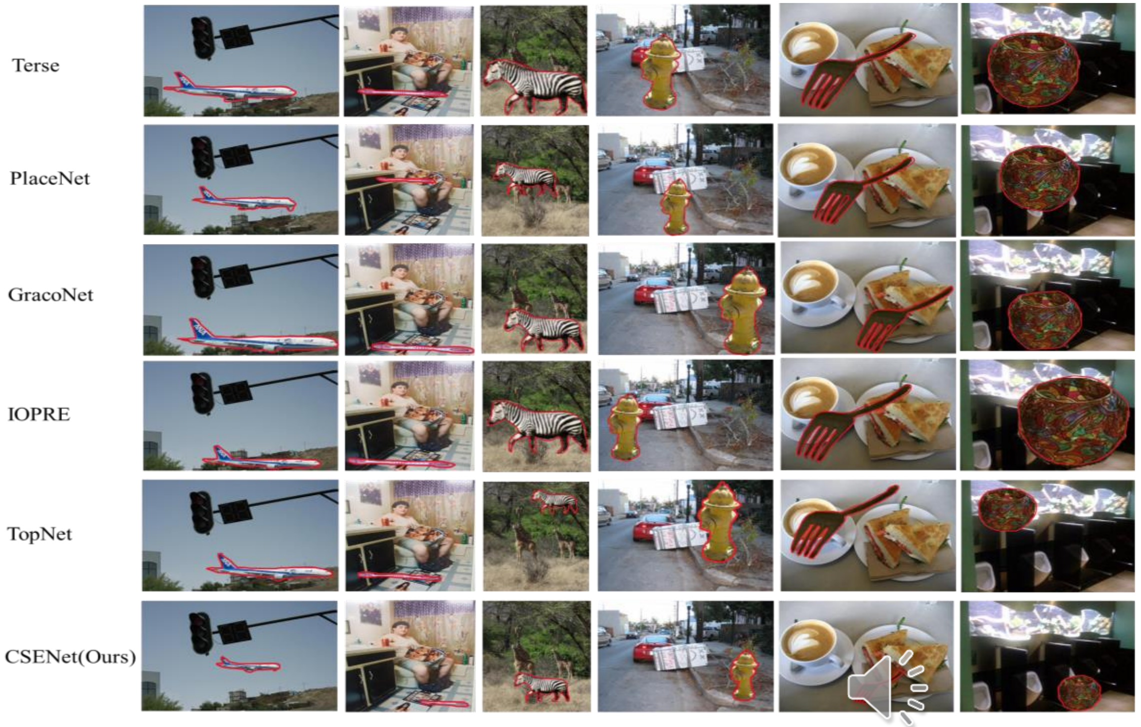
In front of the building on the sidewalk.

Next to the toilet, on the floor.

Beside the sandwich.

In front of the two giraffes.

The boat should be put in the river.



Thanks for Watching and welcome to our poster!

Think before Placement: Common Sense Enhanced Transformer for Object Placement

Yaxuan Qin, Jiayu Xu, Ruiping Wang, Xilin Chen

Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
{qinyaxuan21s, xujiayu22s, wangruiping, xlchen}@ict.ac.cn

Fri 4 Oct 10:30 a.m. CEST — 12:30 p.m. CEST