國立臺灣大學 National Taiwan University

Qualcomm

Scene Graph Expansion for Semantics-Guided Image Outpainting

Chiao-An Yang, Cheng-Yo Tan, Wan-Cyuan Fan
Cheng-Fu Yang, Meng-Lin Wu, Yu-Chiang Frank Wang

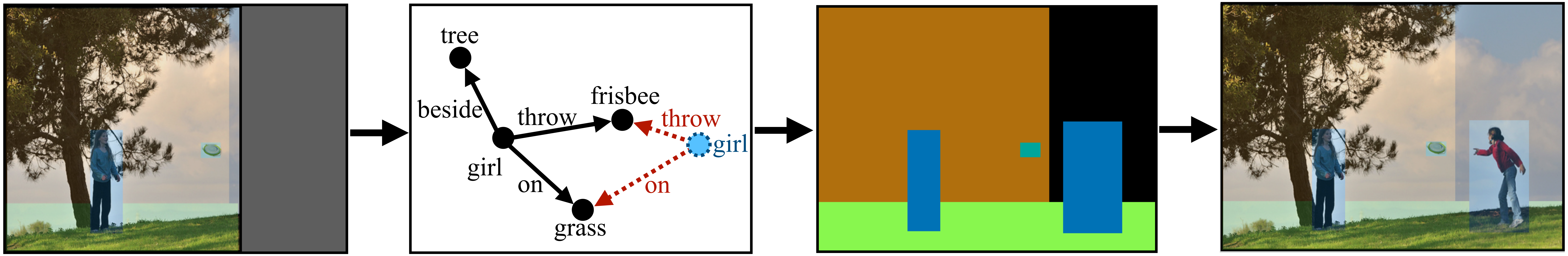


Image Outpainting

Generate the visual context of an image beyond its given boundary.

Input

Output

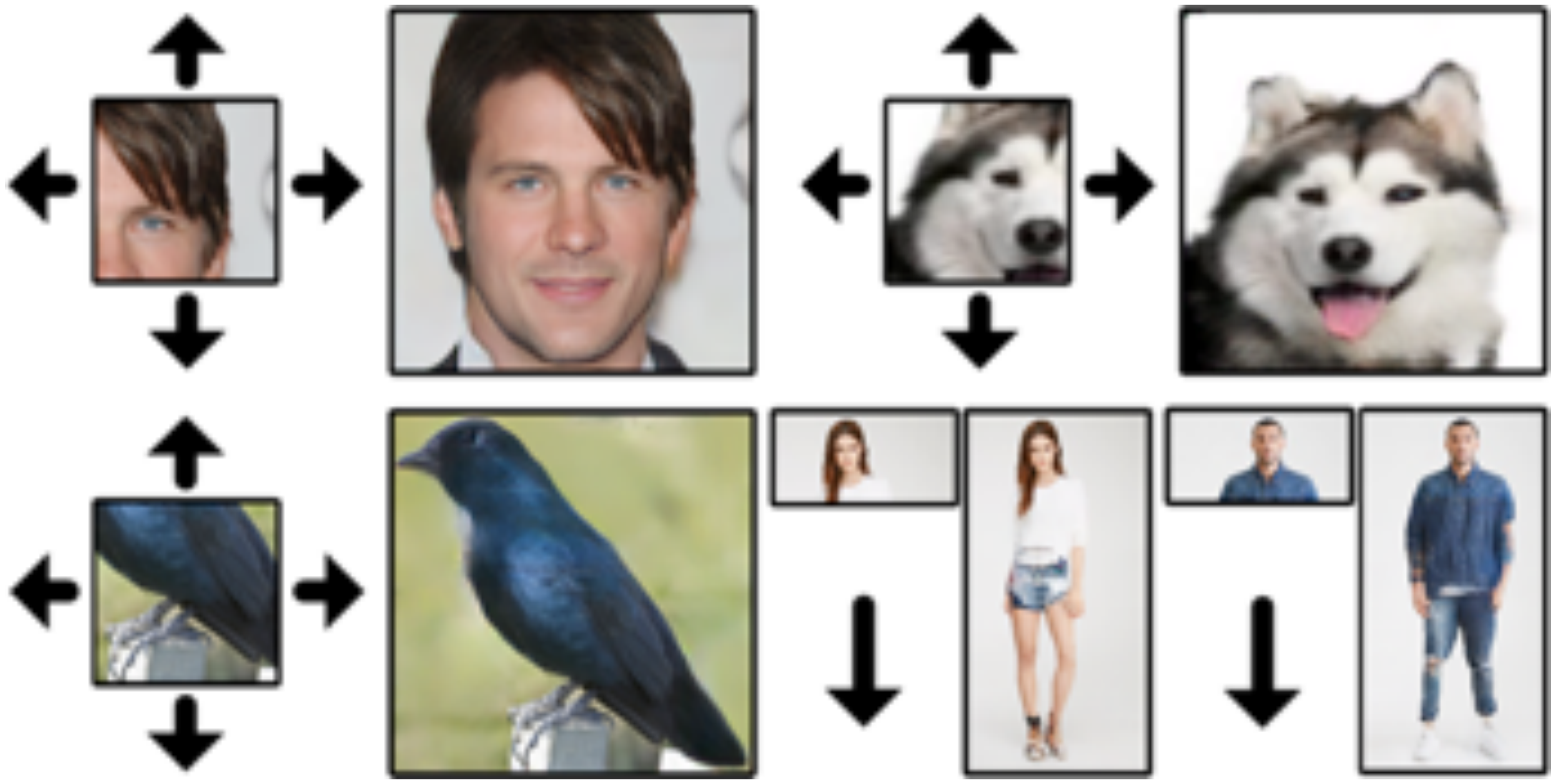
Observation

- Most previous approaches focus on…

1. extending the surrounding texture
   eg. sceneries

[1]

2. completing the fractional objects
   eg. faces, birds, clothes



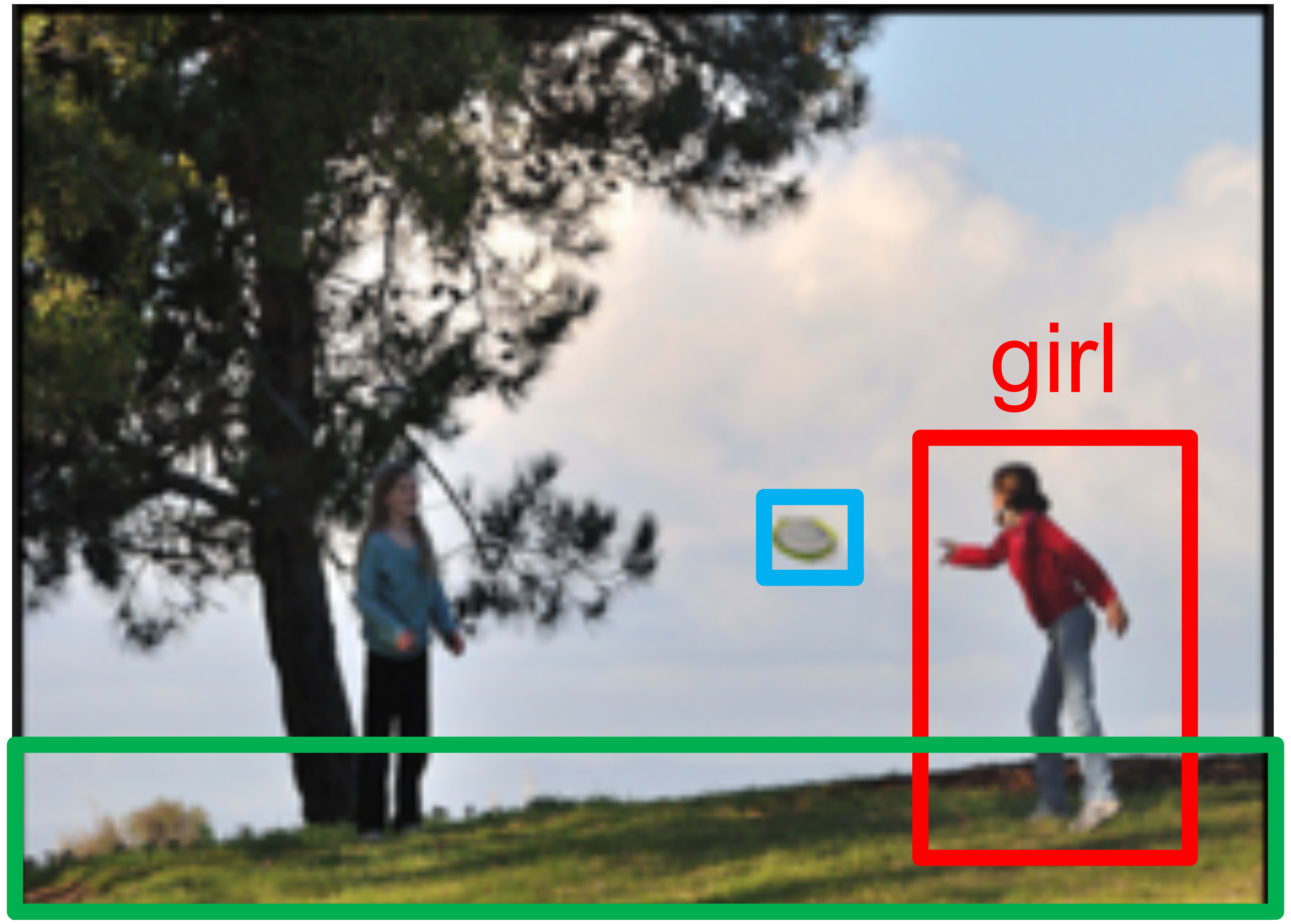
[2]

[1] *Zongxin Yang, Jian Dong, Ping Liu, Yi Yang, and Shuicheng Yan.* Very long natural scenery image prediction by outpainting. ICCV, 2019.
[2] *Yi Wang, Xin Tao, Xiaoyong Shen, and Jiaya Jia.* Wide-context semantic image extrapolation. CVPR, 2019

A Toy Example

Motivation

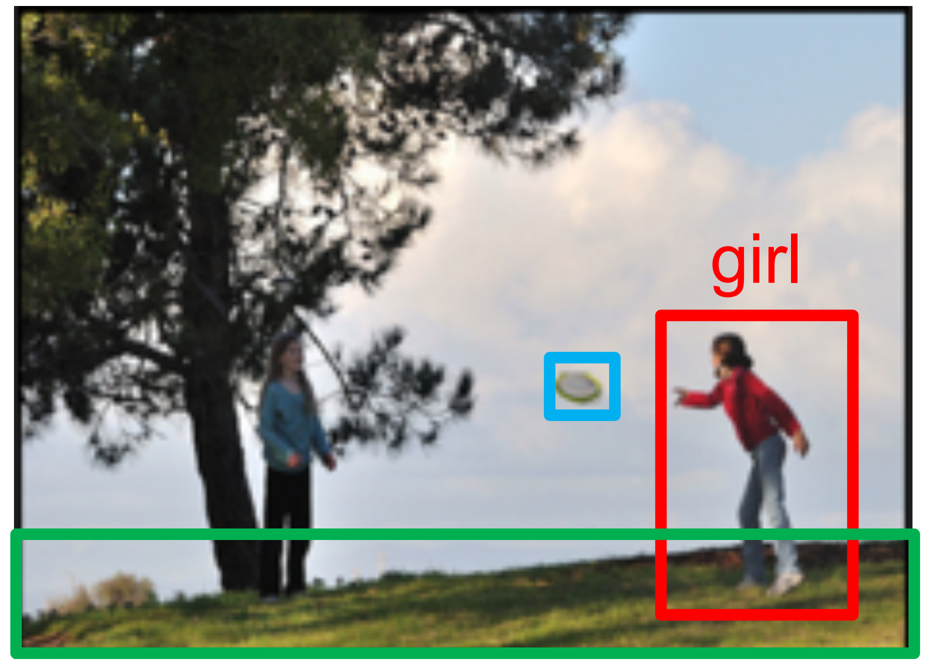
- To generate novel **object**(s) with reasonable **relationships**

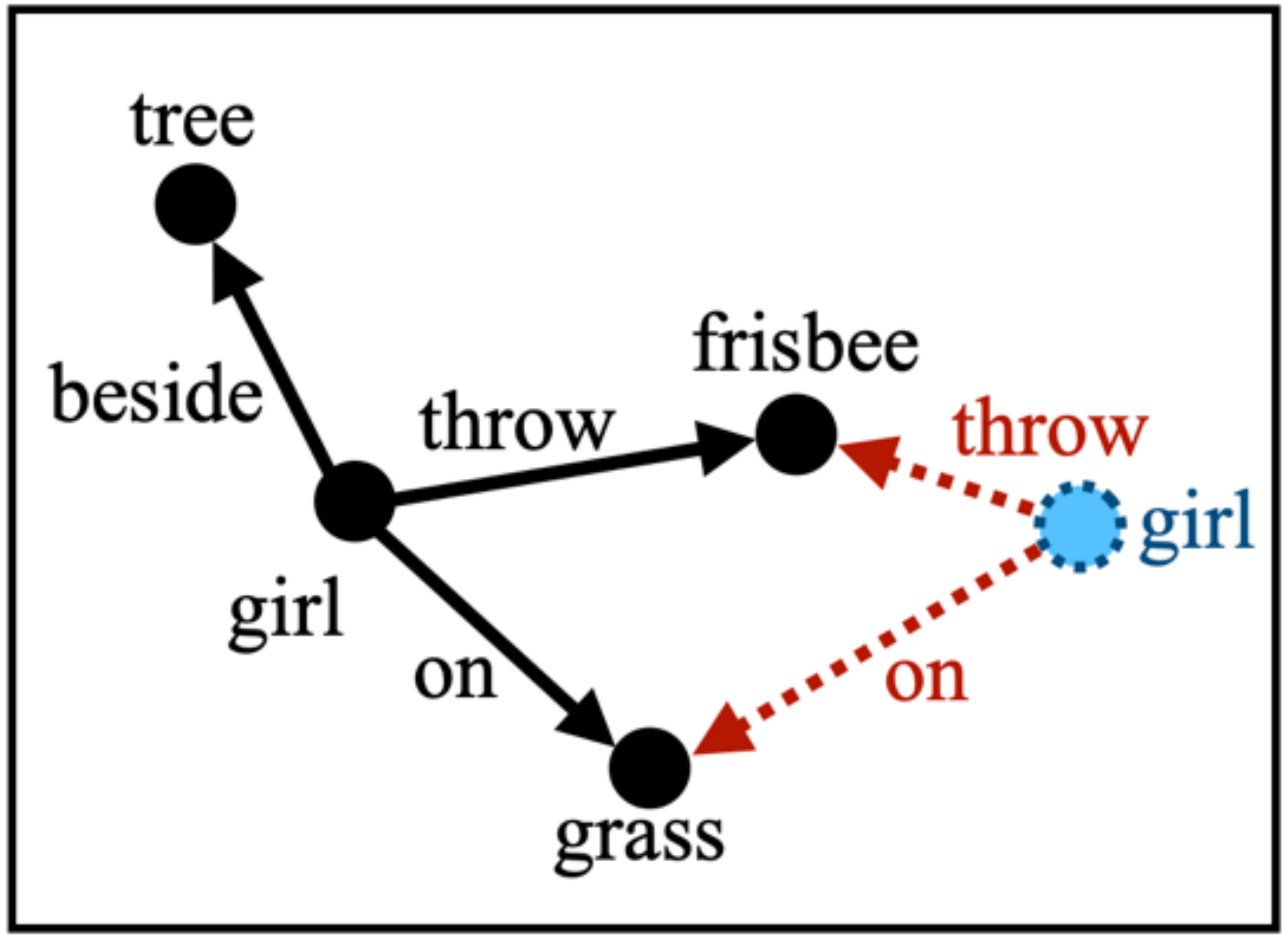


girl throw frisbee
girl on grass

Scene Graphs

- To analyze both **objects** and **relationships**, **scene graph** is a desirable data representation.

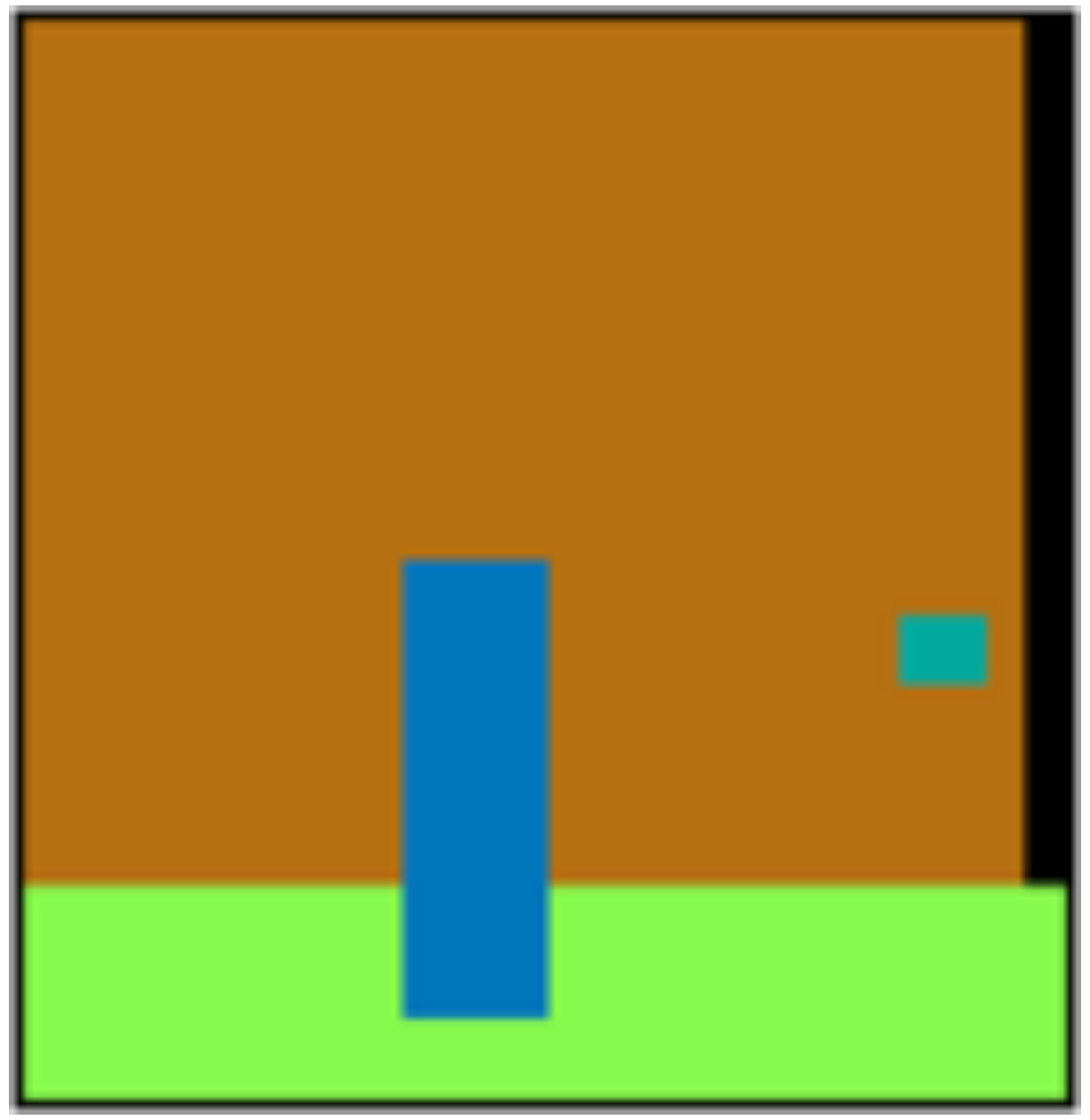




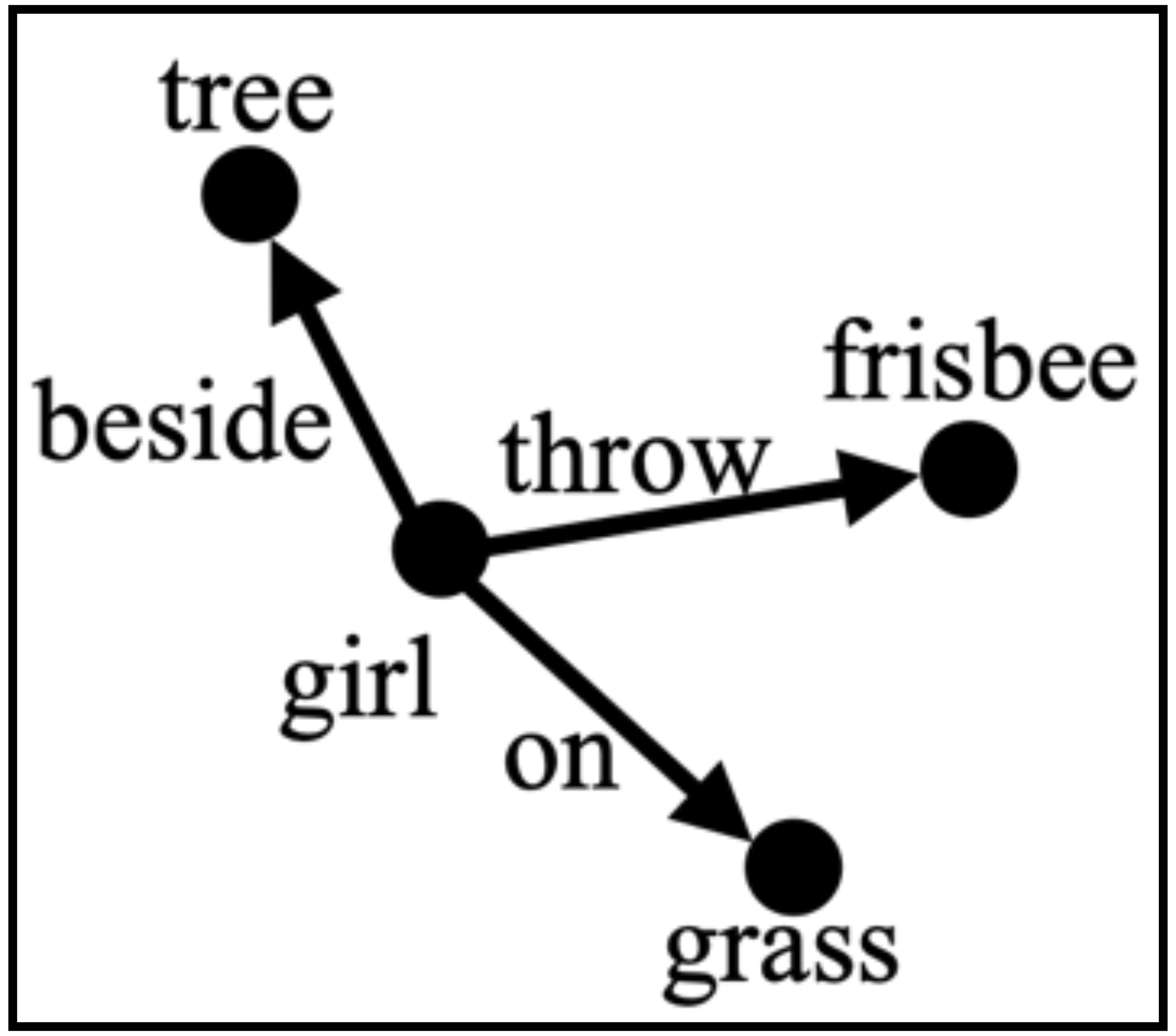
Three-level Representation

- A given image can be decomposed into **three** levels of information.

Visual (RGB image)

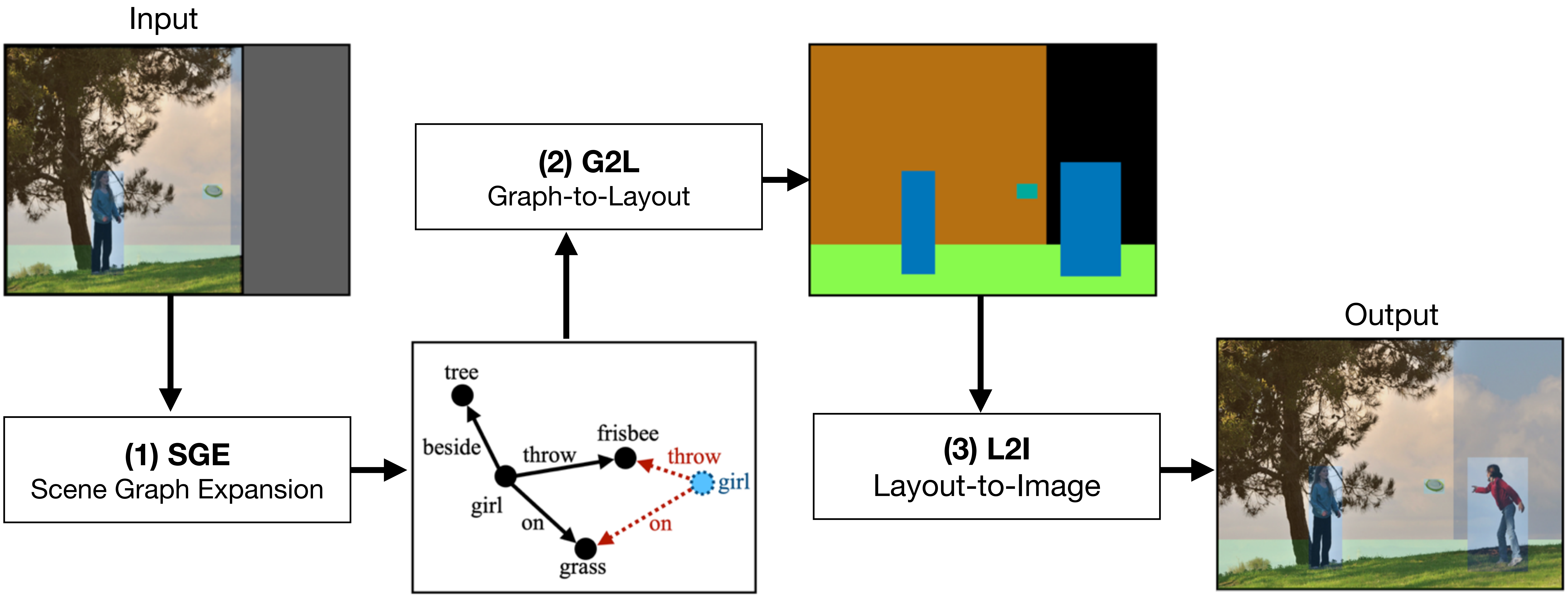


Layout (Bounding box)



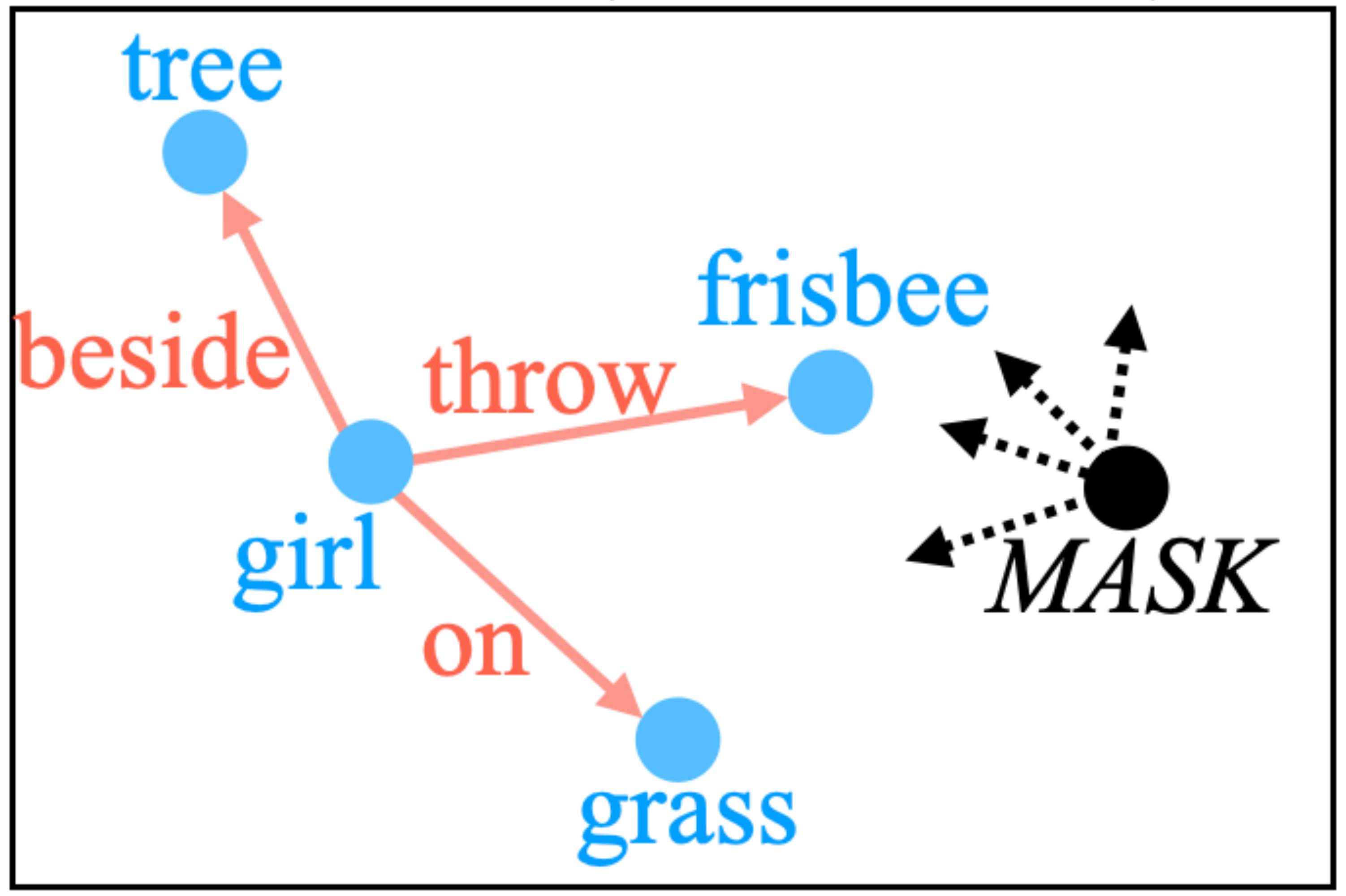
Scene graph

Three-stage Outpainting



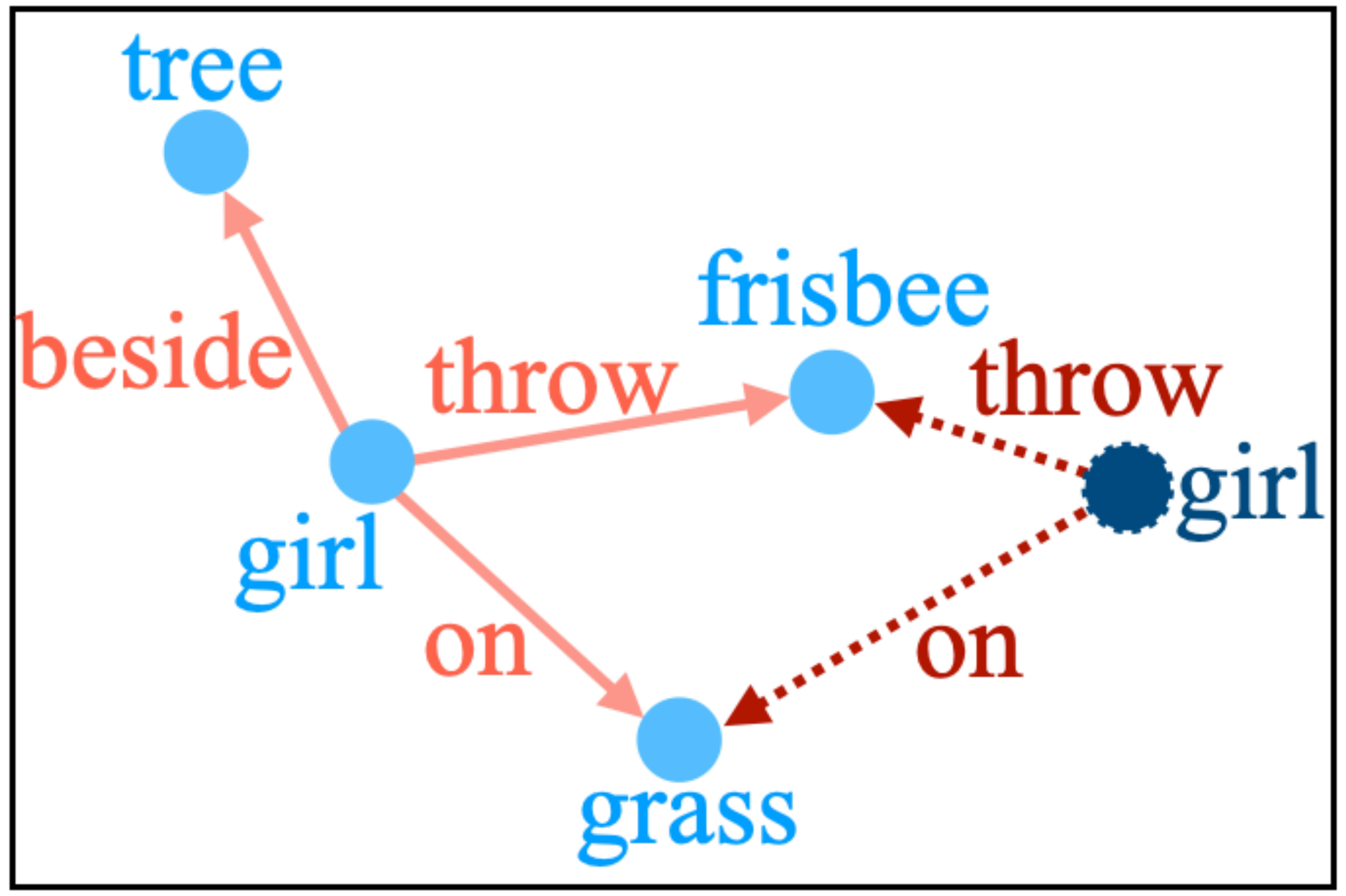
Scene Graph Expansion

$$\mathcal{S}^{in} = (\mathbf{O}^{in}, \mathbf{R}^{in})$$



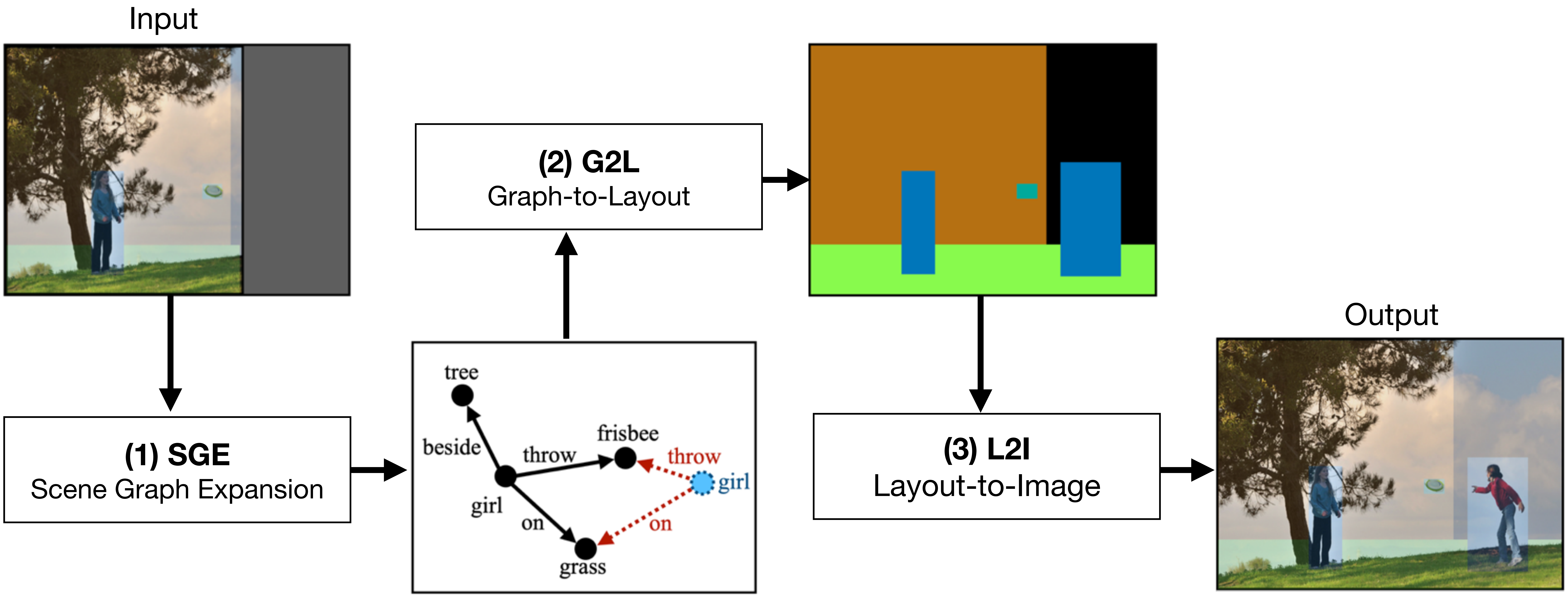
Input SG

$$\mathcal{S}^{op} = (\mathbf{O}^{op}, \mathbf{R}^{op})$$



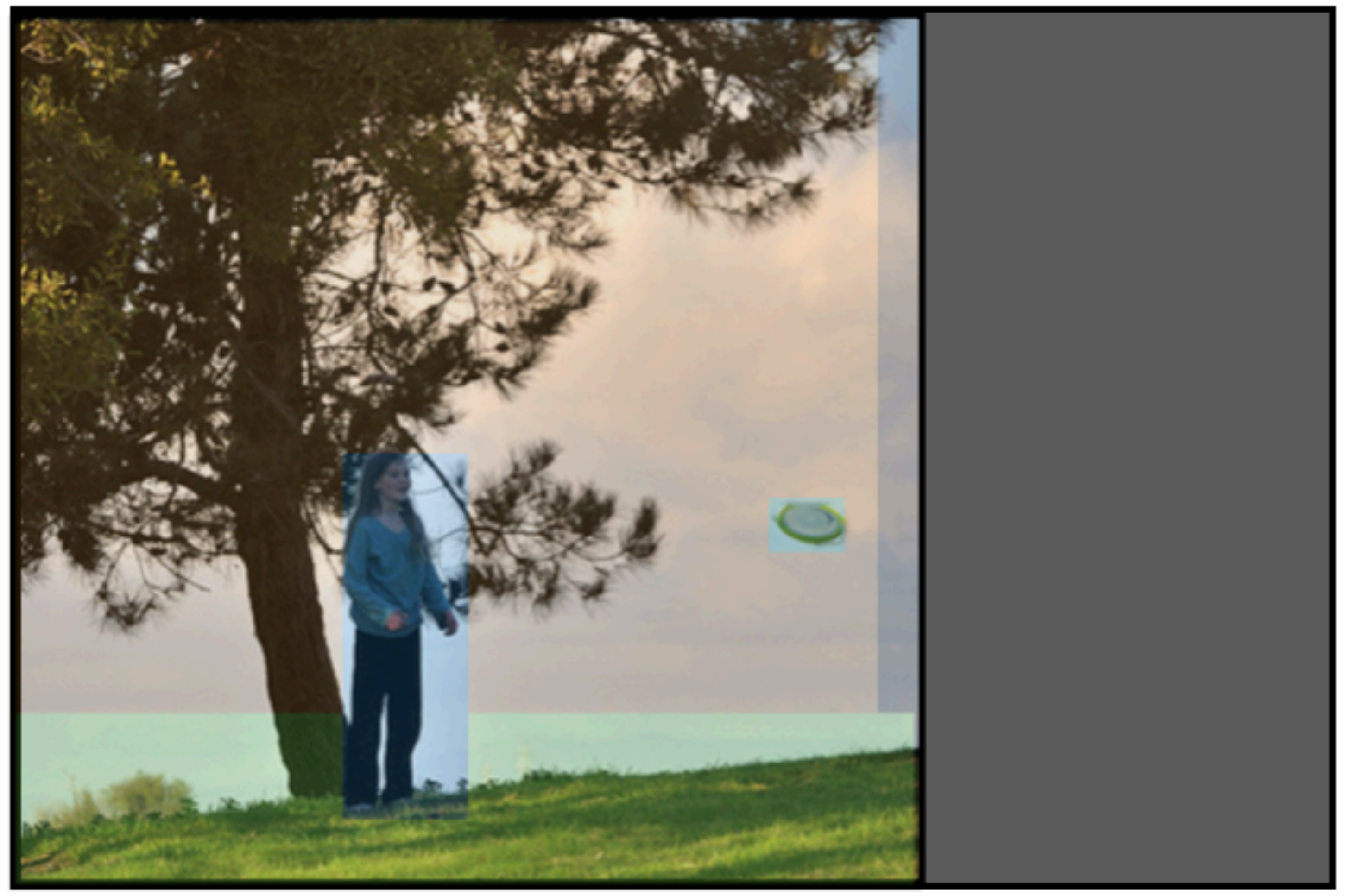
Output SG

Three-stage Outpainting



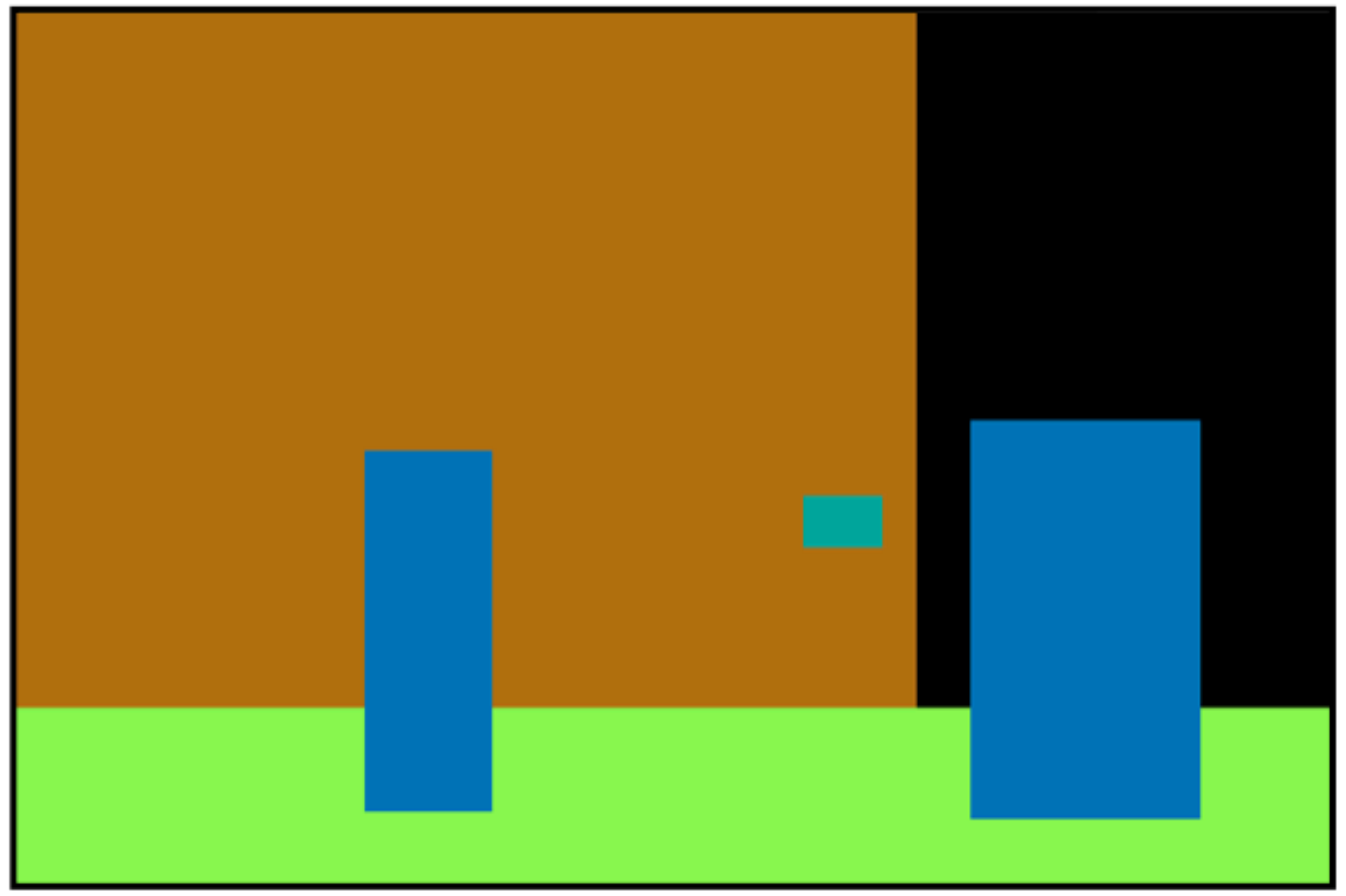
Graph-to-Layout

$$(\mathbf{L}^{in}, \mathbf{I}^{in}) = (\mathbf{B}^{in}, \mathbf{D}^{in}, \mathbf{I}^{in})$$



Input layout, image

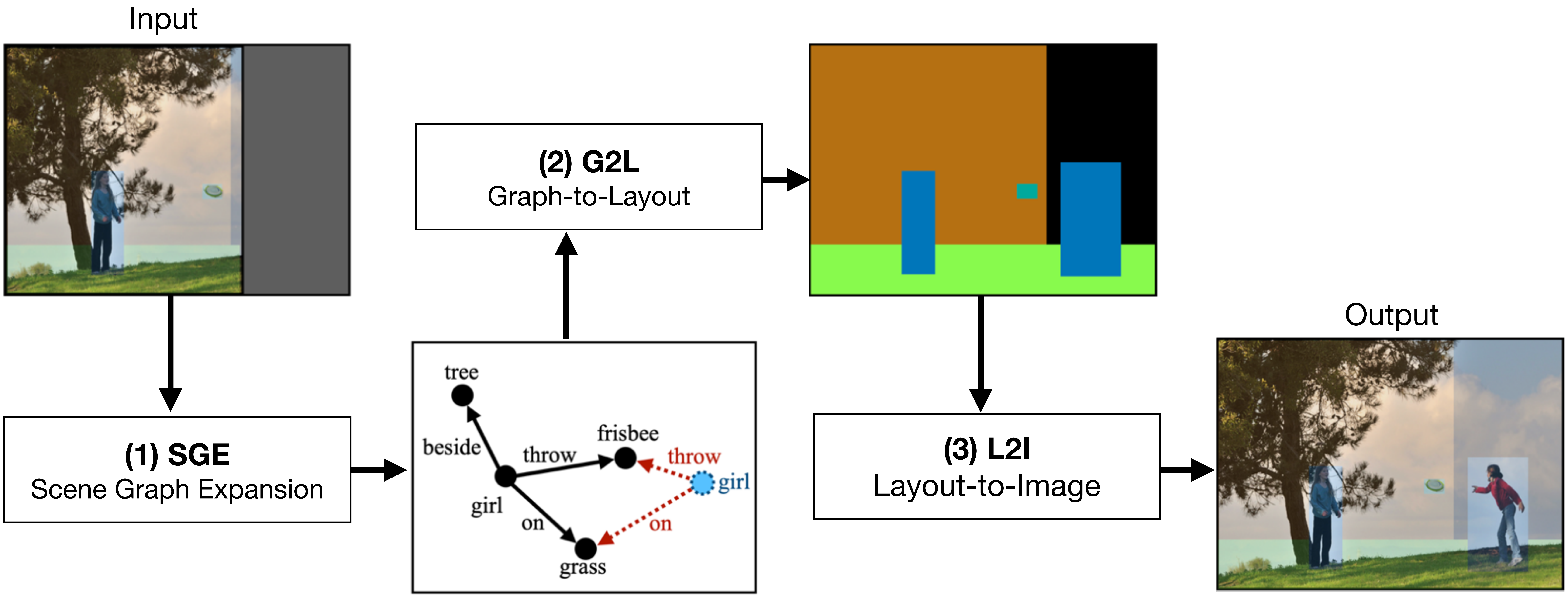
$$\mathbf{L}^{op} = (\mathbf{B}^{op}, \mathbf{D}^{op})$$



Output layout

$$\begin{cases} B_i & = \text{bounding box of object\_i} \\ D_{ij} & = \text{bounding box displacement between object\_i and object\_j} \end{cases}$$

Three-stage Outpainting



Approach

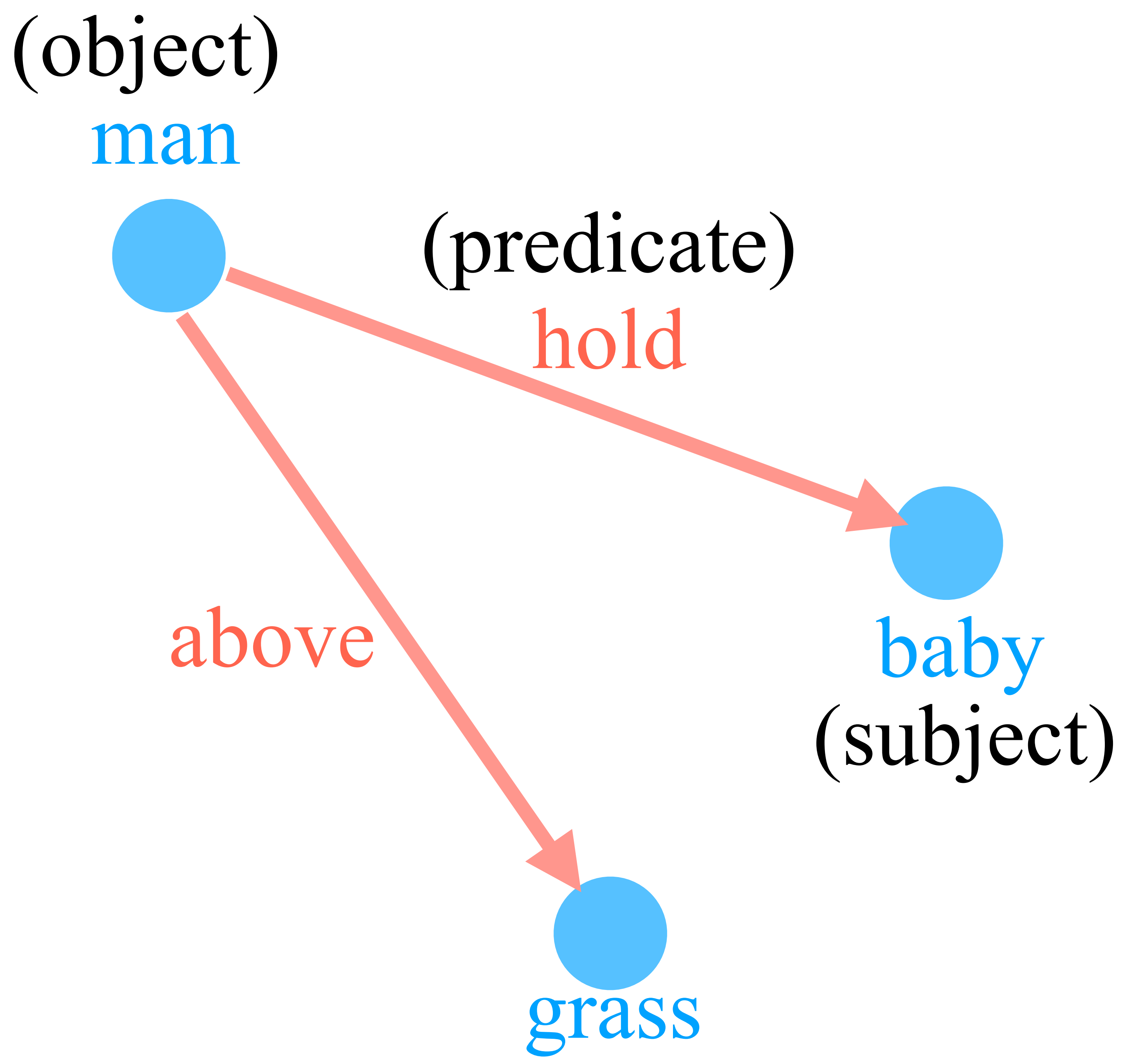
- Transformer-based architecture
  - **SGT:** Scene Graph Transformer
- Semantic-guided Image Outpainting
  - **SGE**: Scene Graph Expansion
  - **G2L**: Graph to Layout
  - **L2I**: Layout to Image

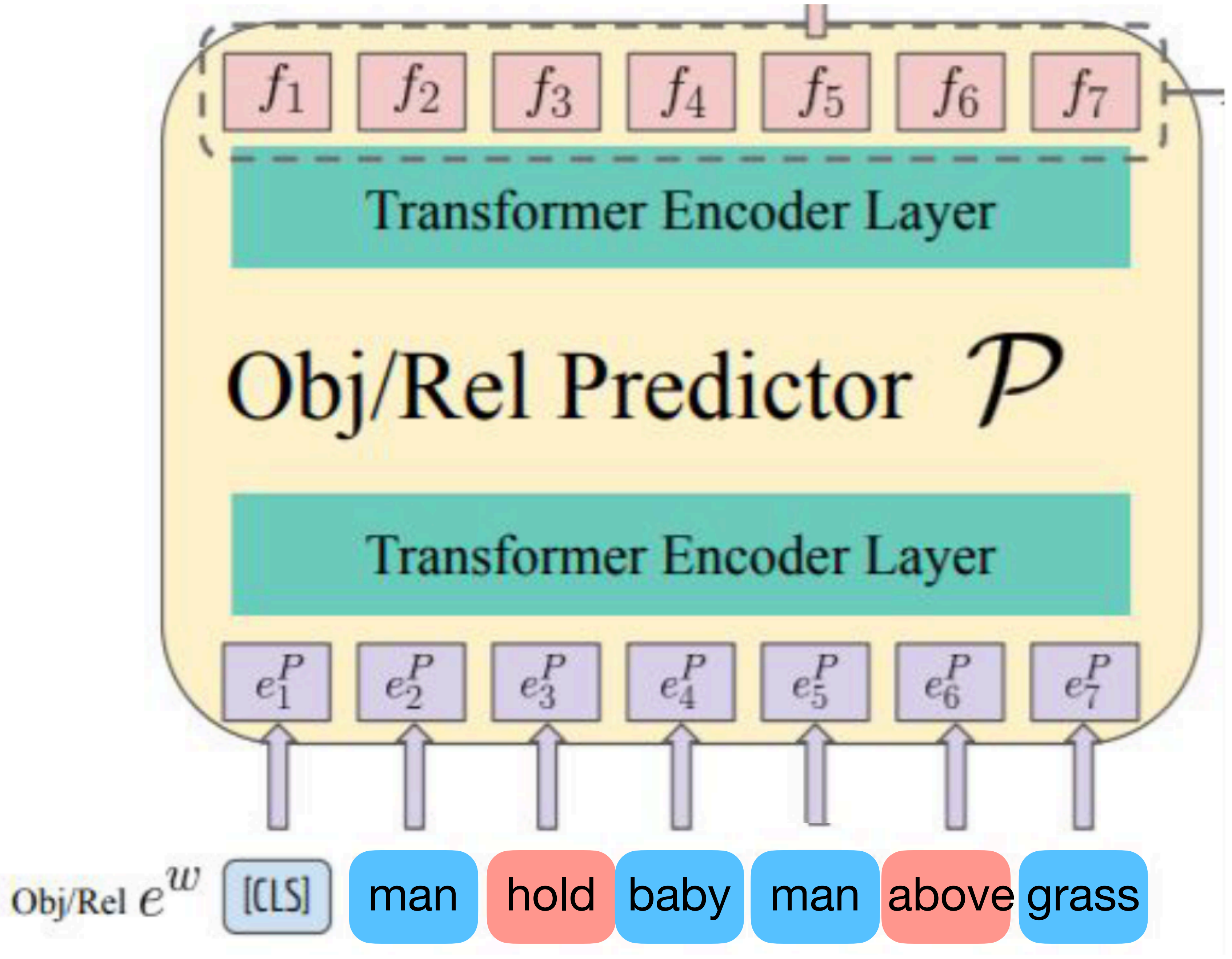
Approach (1)

- Transformer-based architecture
  - **SGT:** Scene Graph Transformer
- Semantic-guided Image Outpainting
  - **SGE**: Scene Graph Expansion
  - **G2L**: Graph to Layout
  - **L2I**: Layout to Image

The Problems of Standard Transformers

- Previous approaches "flatten" the scene graph as triplets sequence which result in long sequence length for a large scene graph $L = 3E = 3N^2$.

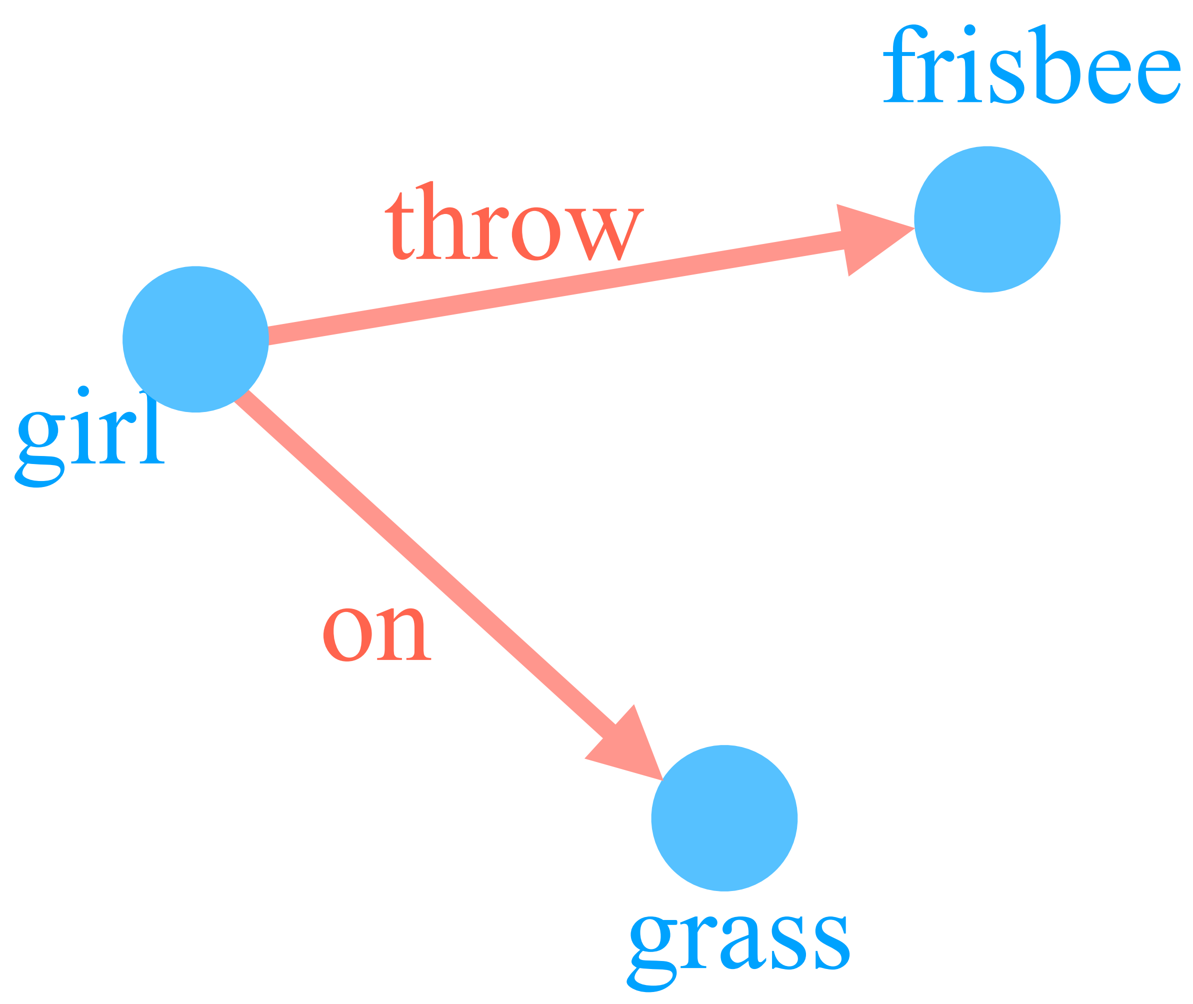




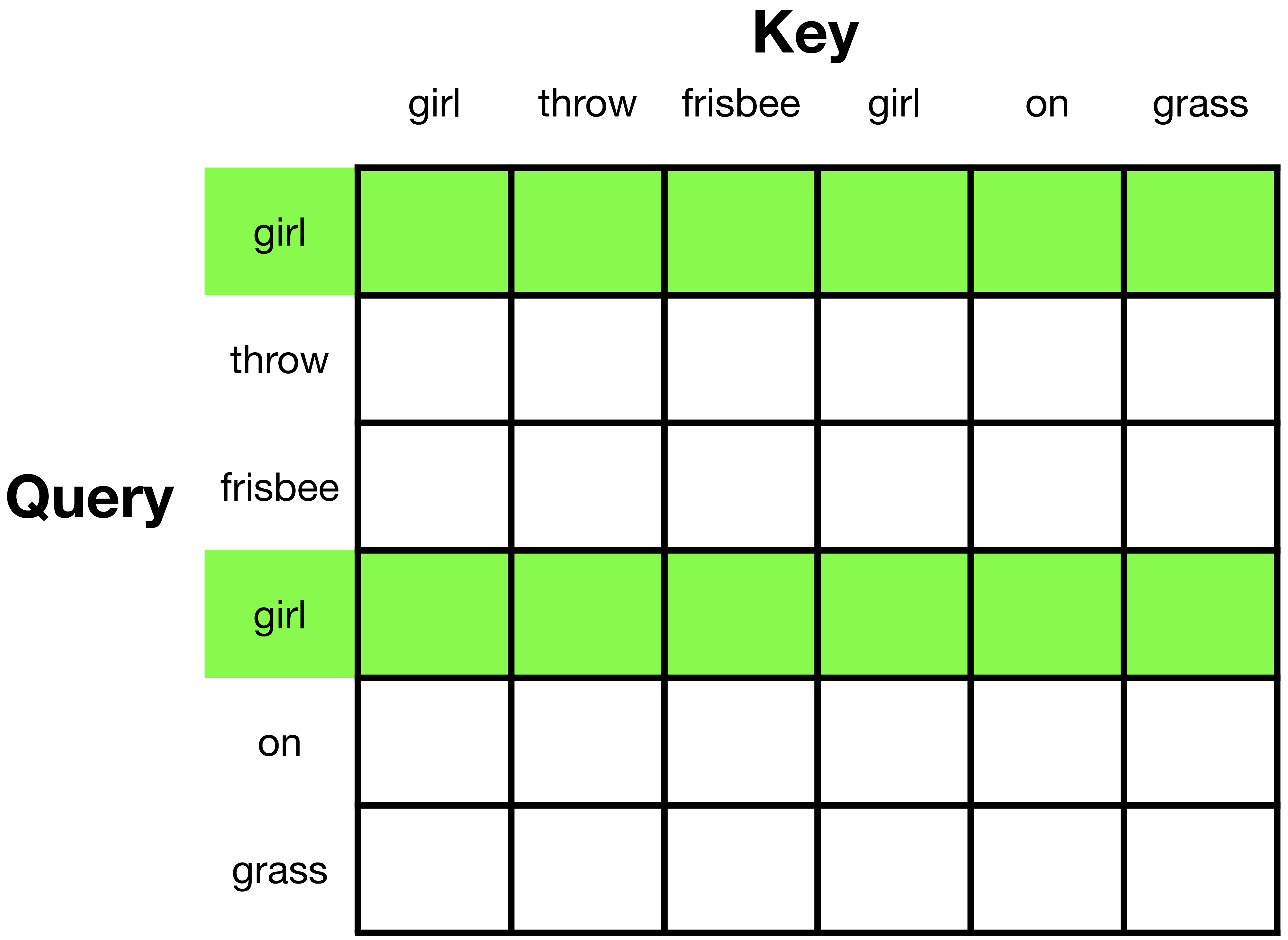
LTNet (Yang et al. CVPR 2021.)

The Problems of Standard Transformers

- It also cause redundant computation since a **single object** with **multiple relationships** will occur in multiple triplets.



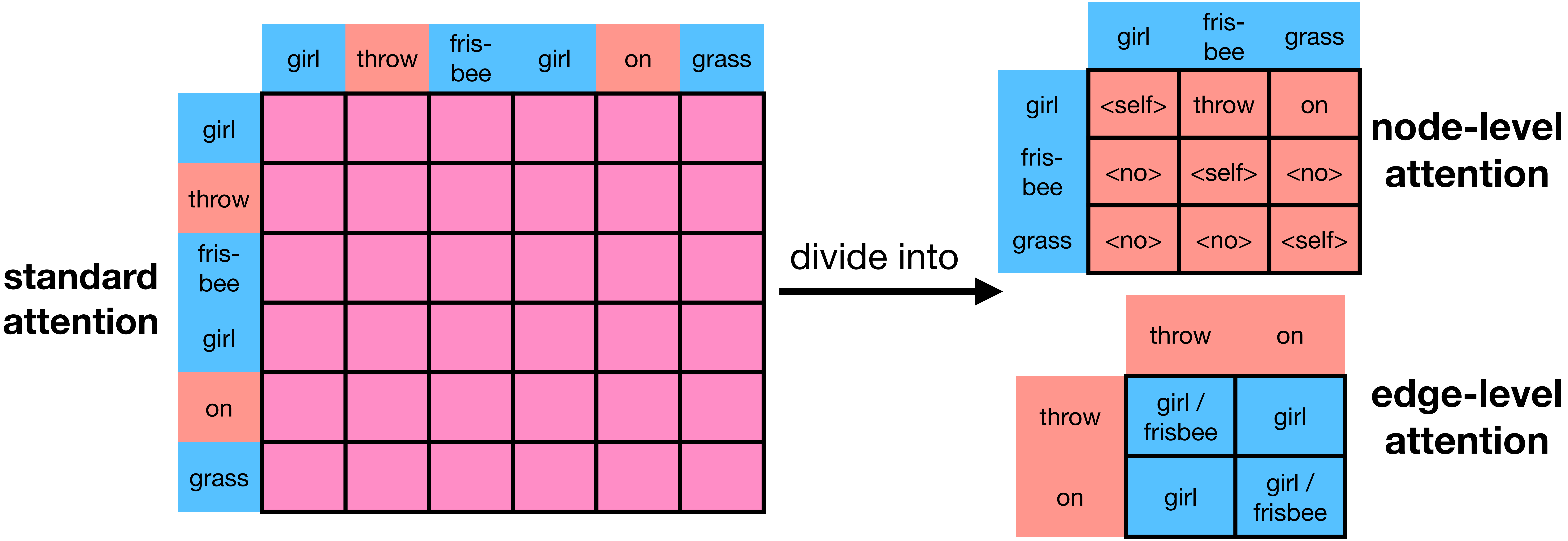
Scene Graph



Redundant Computation of Self-attention

Scene Graph Transformer

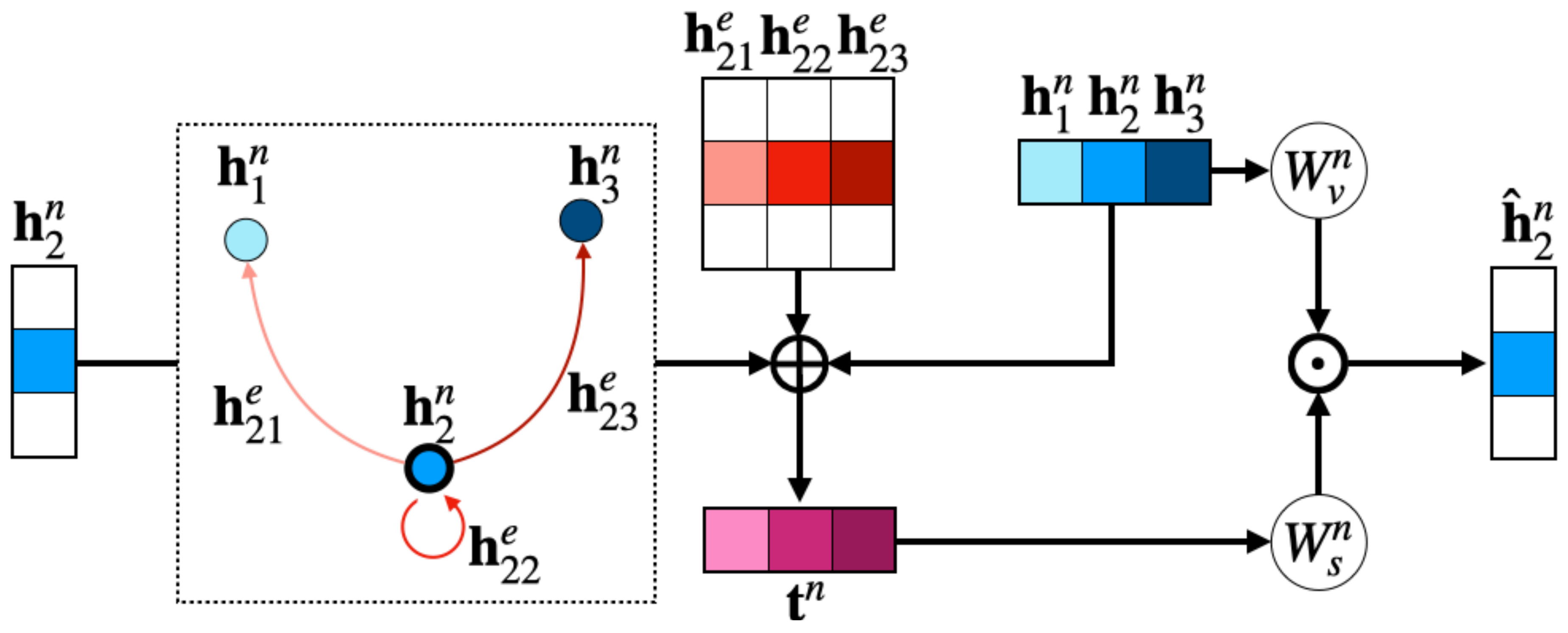
- No long input sequence
- No redundant computation



Node-level attention

- The attention between (node_i, node_j) is dependent on edge_ij.

	girl	frisbee	grass
girl	<self>	throw	on
frisbee	<no>	<self>	<no>
grass	<no>	<no>	<self>

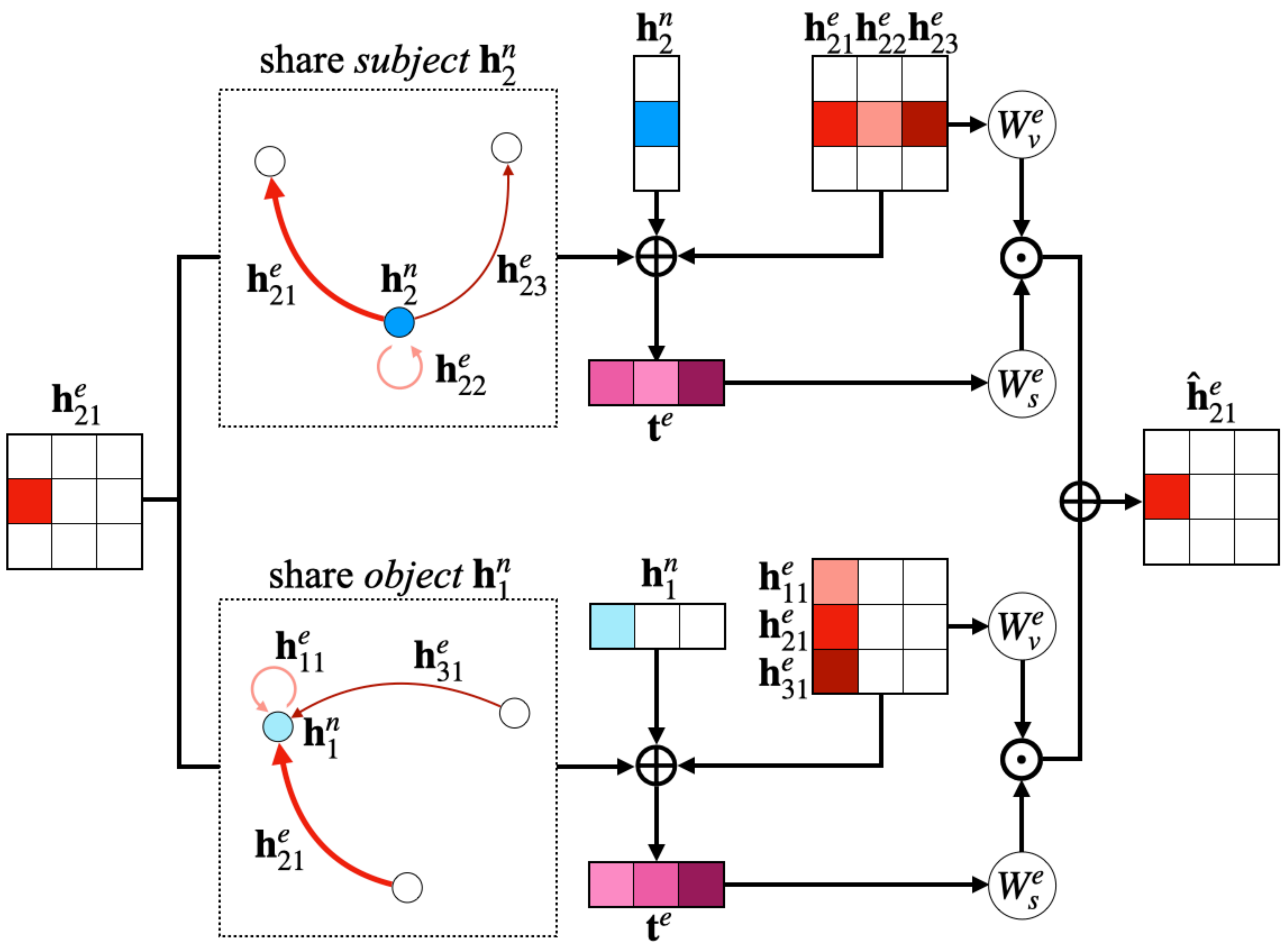


$$\begin{cases} \textbf{attention}(h_2^n, h_1^n) \leftarrow h_{21}^e \\ \textbf{attention}(h_2^n, h_2^n) \leftarrow h_{22}^e \\ \textbf{attention}(h_2^n, h_3^n) \leftarrow h_{23}^e \end{cases}$$

Edge-level attention

- The attention (edge_ij, edge_ik) is dependent on object node_i.
  The attention (edge_ij, edge_kj) is dependent on subject node_j.

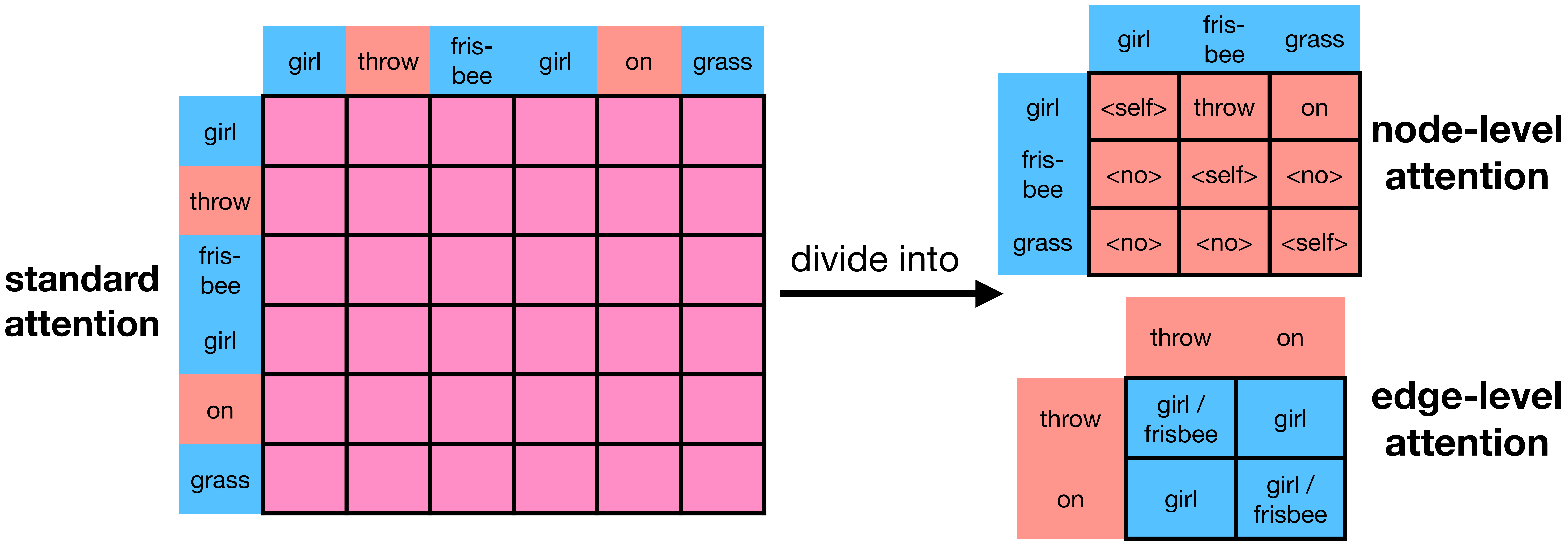
	[girl] **throw** [frisbee]	[girl] **on** [grass]
[girl] **throw** [frisbee]	girl / frisbee	girl
[girl] **on** [grass]	girl	girl / frisbee



$$\left\{\begin{array}{l}\textbf{attention}(h^e_{21}, h^e_{21}) \leftarrow h^n_2, h^n_1 \\ \textbf{attention}(h^e_{21}, h^e_{22}) \leftarrow h^n_2 \\ \textbf{attention}(h^e_{21}, h^e_{23}) \leftarrow h^n_2 \\ \textbf{attention}(h^e_{21}, h^e_{11}) \leftarrow h^n_1 \\ \textbf{attention}(h^e_{21}, h^e_{31}) \leftarrow h^n_1 \end{array}\right.$$

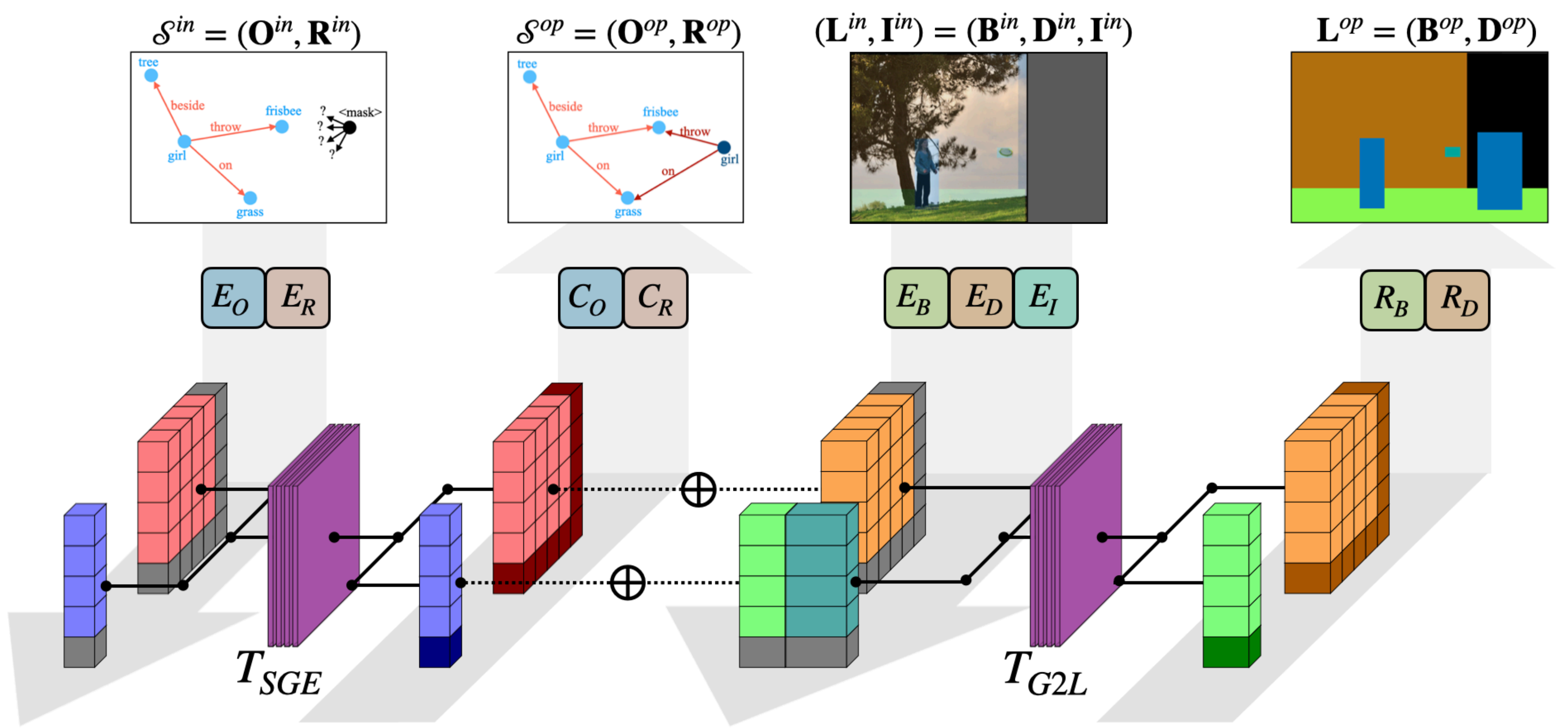
Scene Graph Transformer

- No long input sequence $L = 3E = 3N^2 \Rightarrow L' = N + N^2$
- No redundant computation $\Rightarrow$ Each node and edge appears once.

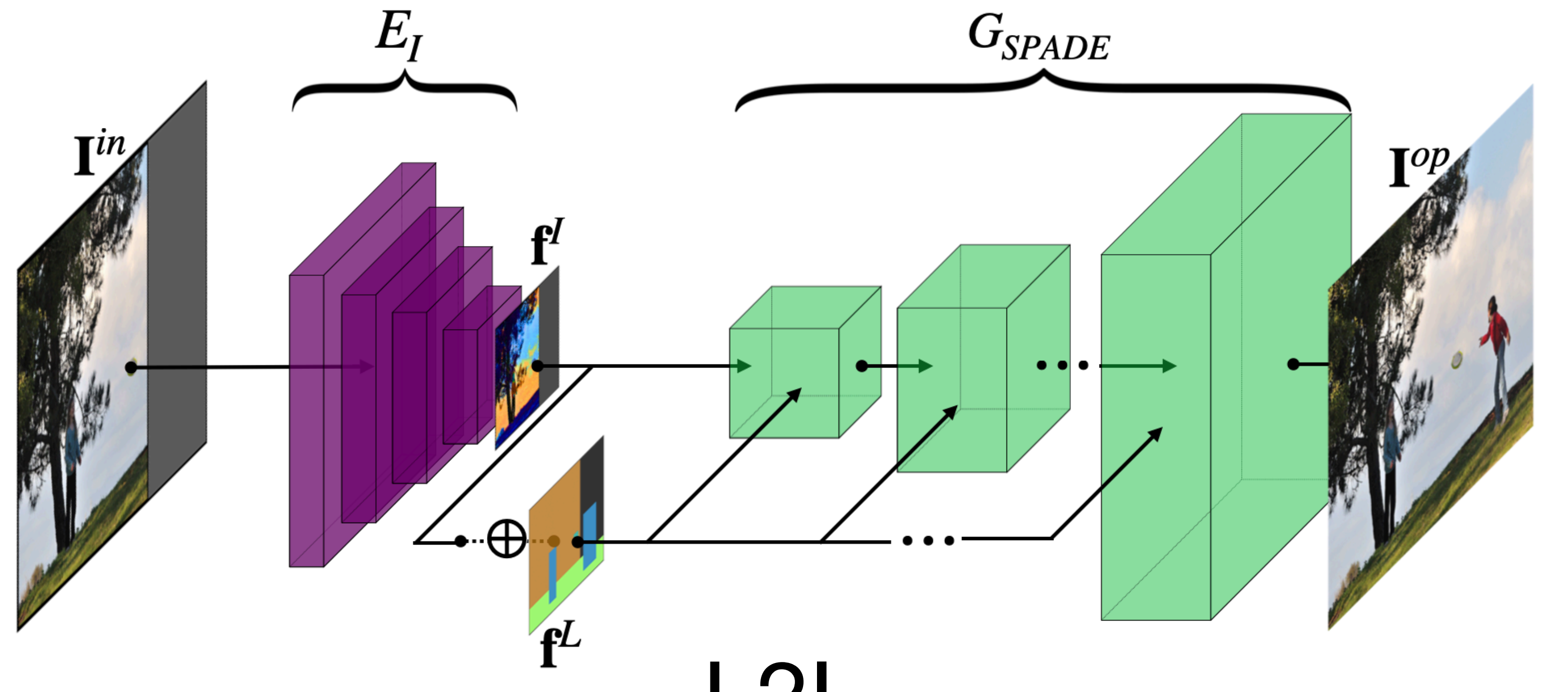


Approach (2)

- Transformer-based architecture
  - **SGT:** Scene Graph Transformer
- Semantic-guided Image Outpainting
  - **SGE**: Scene Graph Expansion
  - **G2L**: Graph to Layout
  - **L2I**: Layout to Image

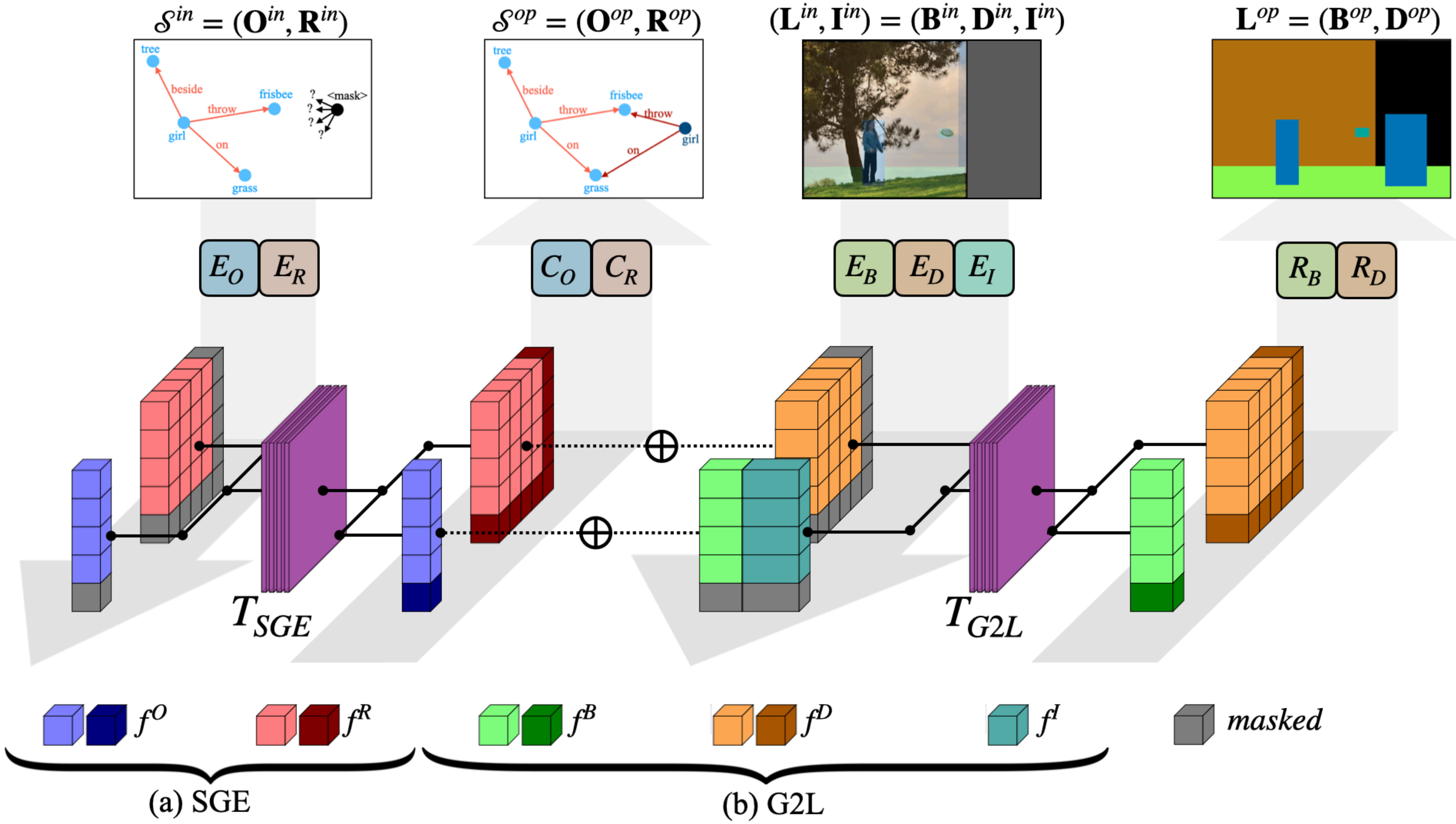


SGE & G2L



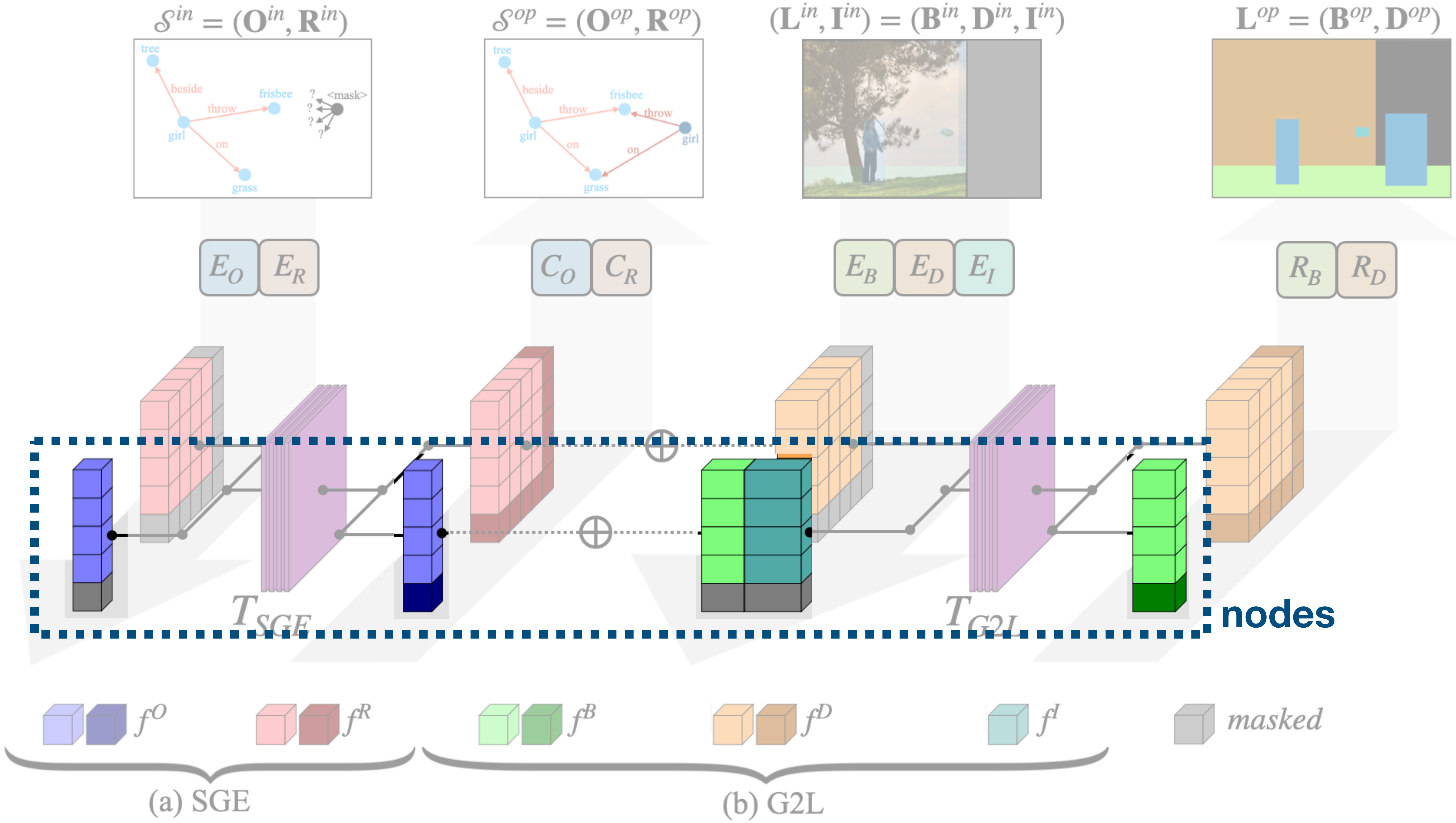
L2I

SGE & G2L



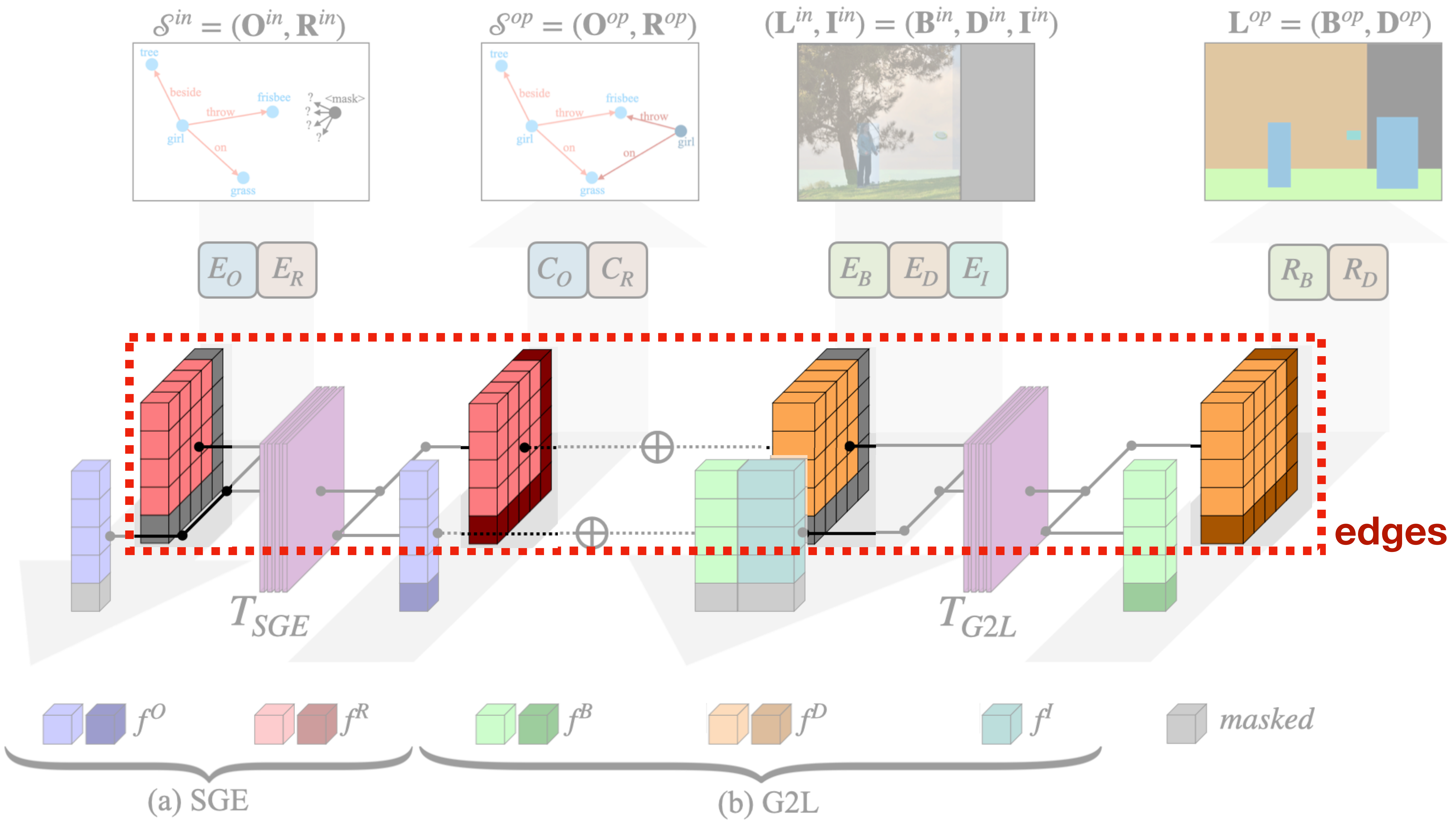
- Both T_{SGE} and T_{G2L} are SG Transformers.
  - node: objects, bboxes…
  - edge: relationships, bbox disparities…

SGE & G2L



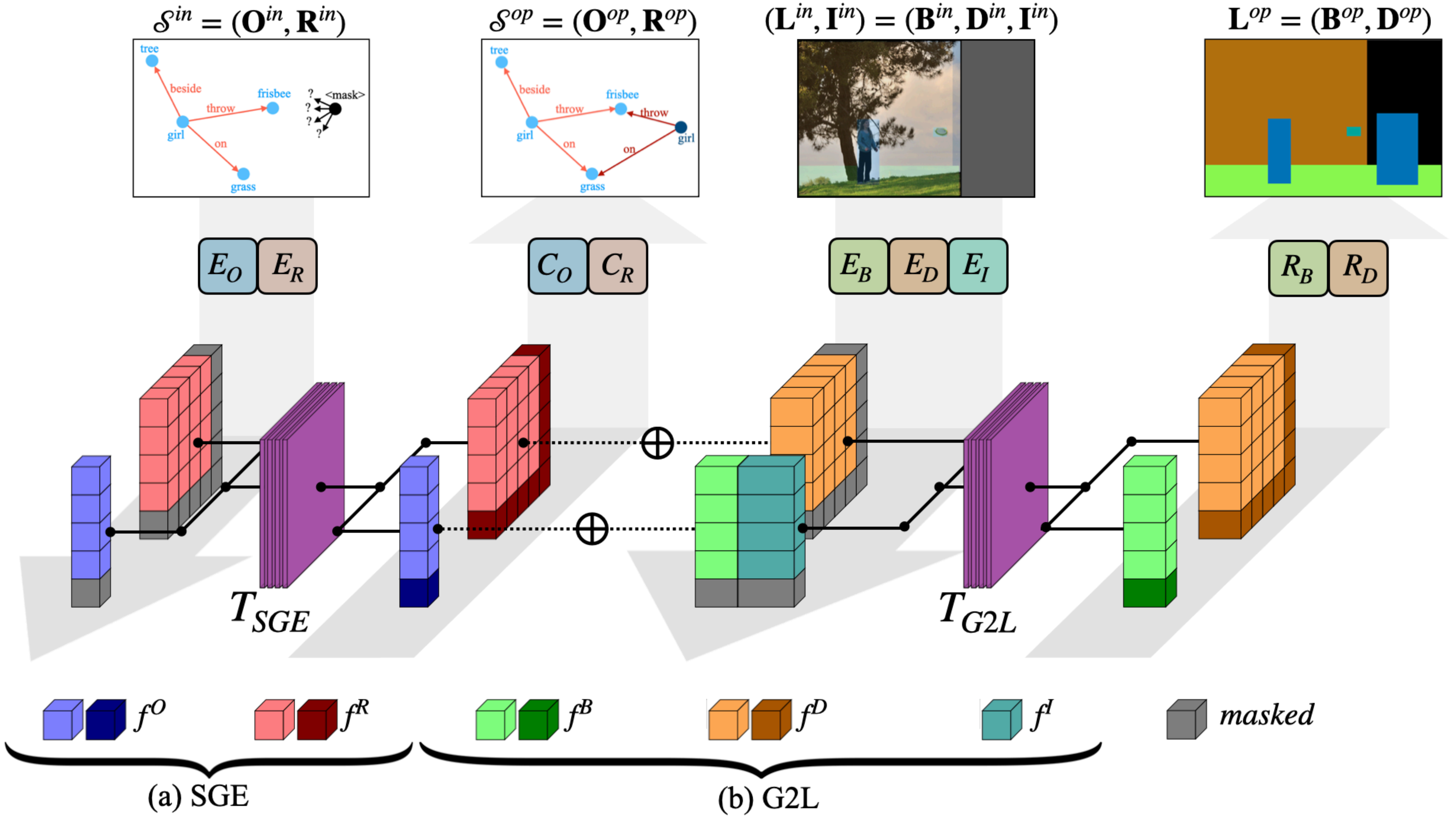
- Both T_{SGE} and T_{G2L} are SG Transformers.
  - node: objects, bboxes…
  - edge: relationships, bbox disparities…

SGE & G2L



- Both T_{SGE} and T_{G2L} are SG Transformers.
  - node: objects, bboxes…
  - edge: relationships, bbox disparities…

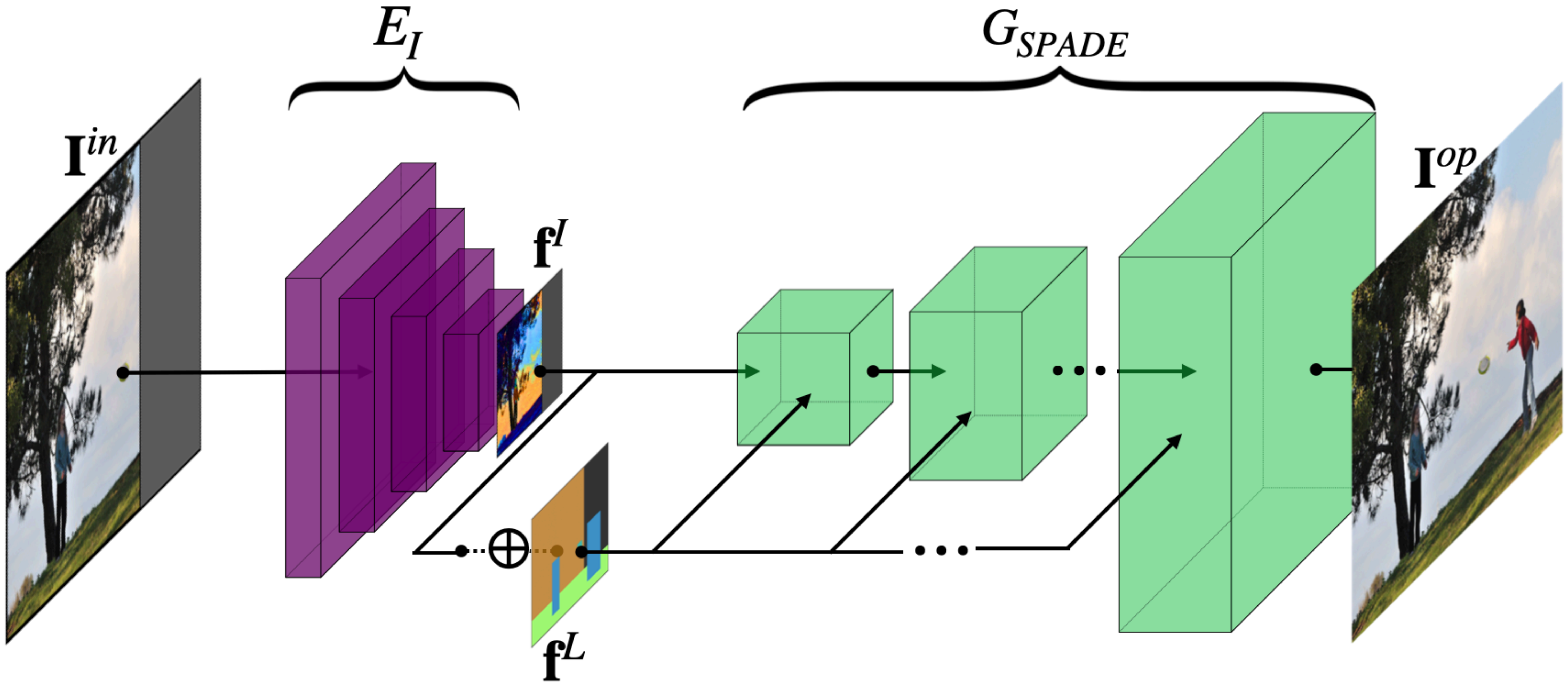
SGE & G2L



- Both T_{SGE} and T_{G2L} are SG Transformers.
  - node: objects, bboxes…
  - edge: relationships, bbox disparities…

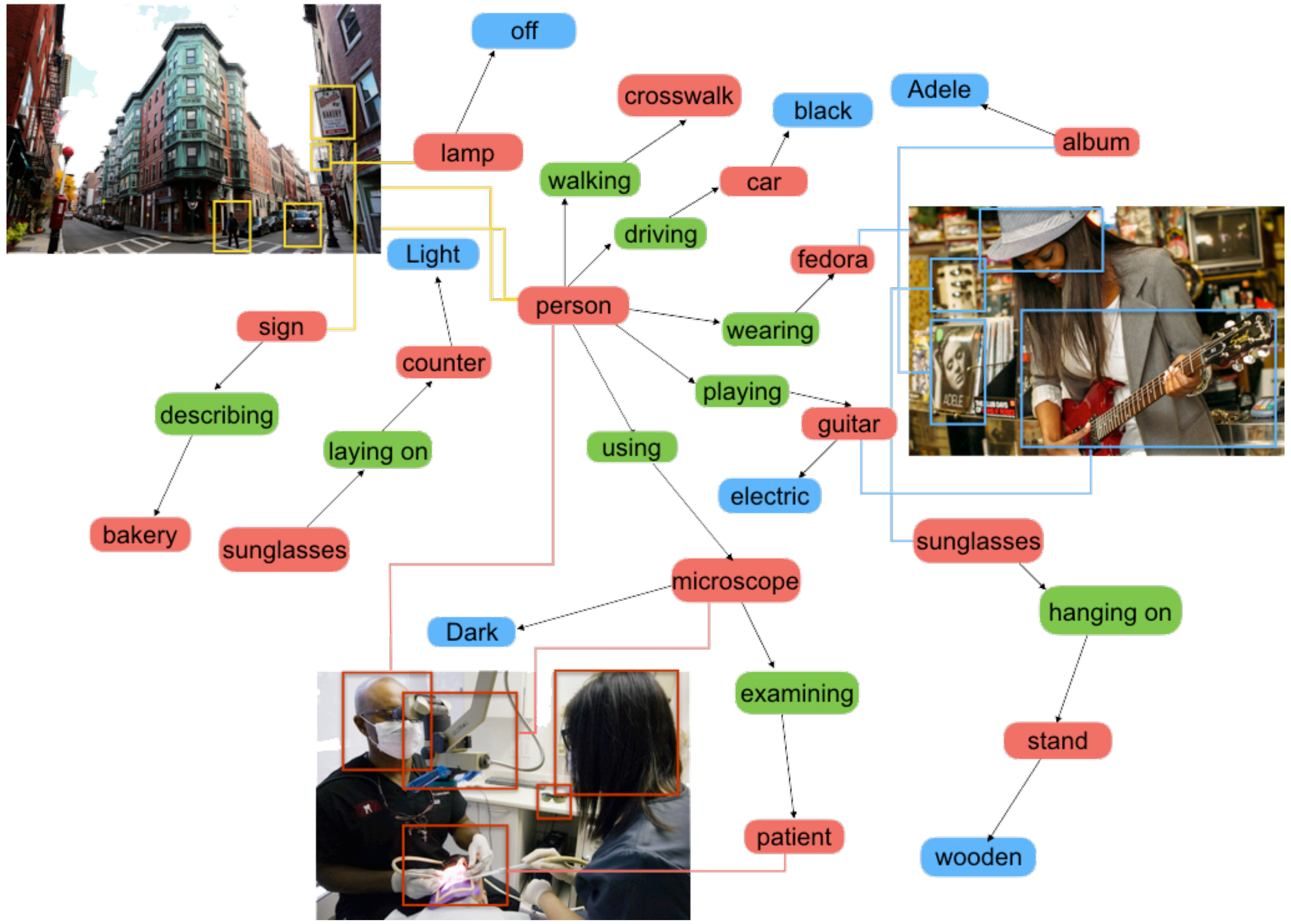
Layout-to-Image

- SPADE-based Generative model [1, 2]
  - semantic map guidance: image f^I + layout f^L

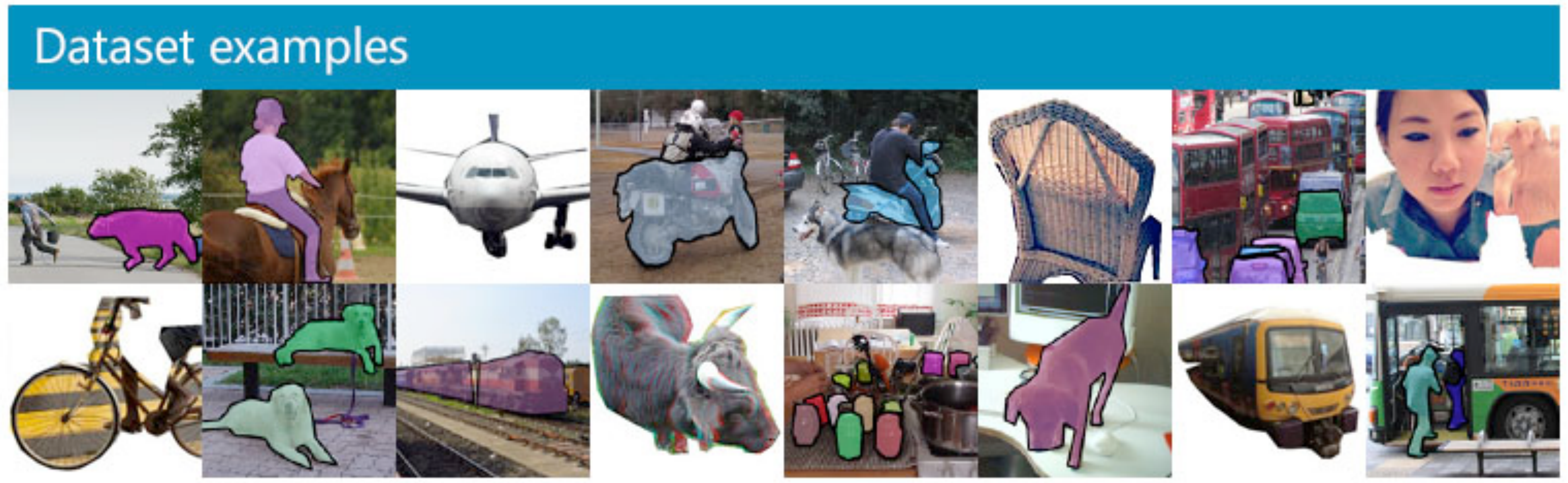


[1] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. CVPR, 2019.
[2] Roei Herzig, Amir Bar, Huijuan Xu, Gal Chechik, Trevor Darrell, and Amir Globerson. Learning canonical representations for scene graph to image generation. ECCV 2020.

Datasets



VG-MSDN



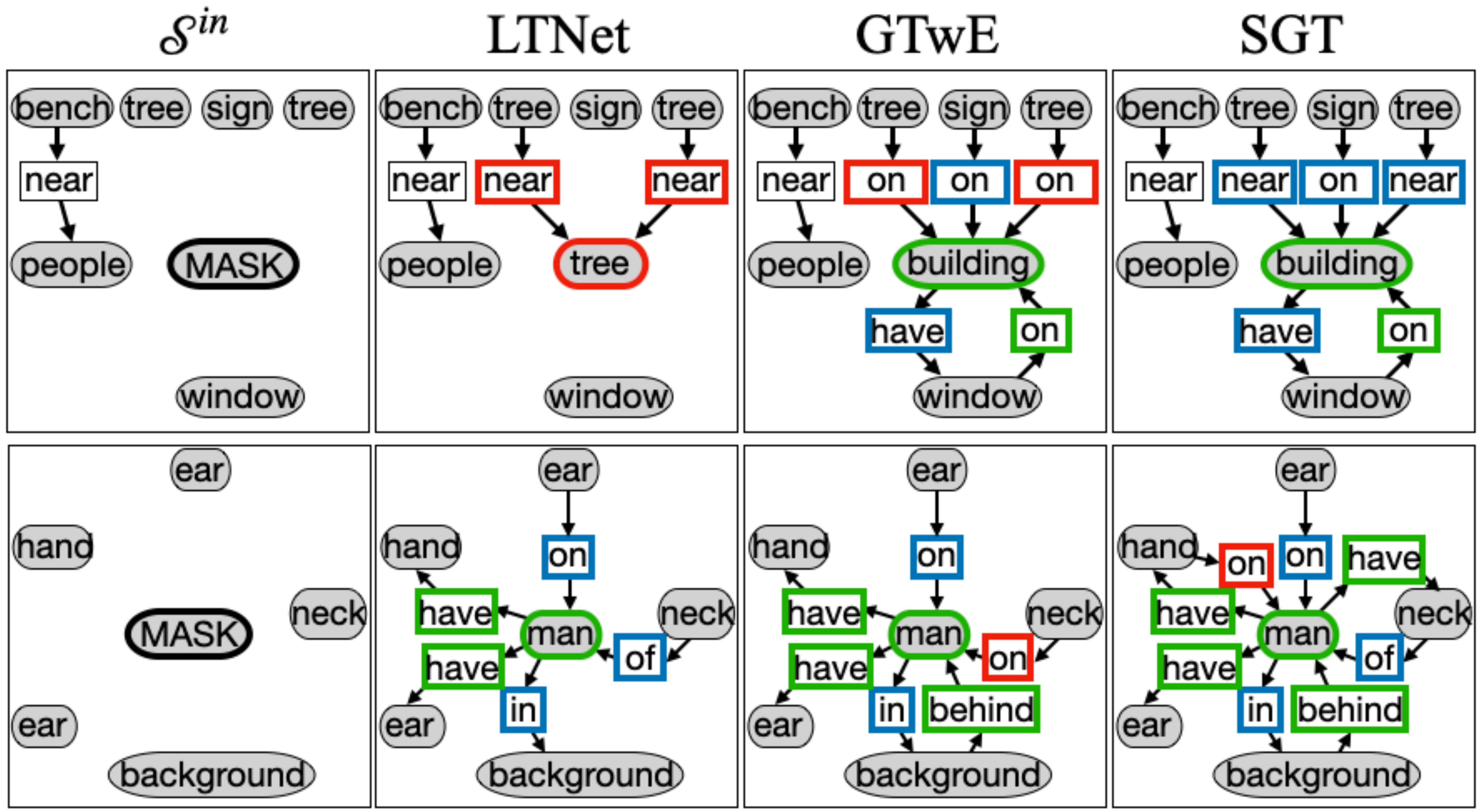
COCO-stuff

Experiments

- The accuracy of our semantics extrapolation
- The quality of our outpainted images

Experiments: SGE

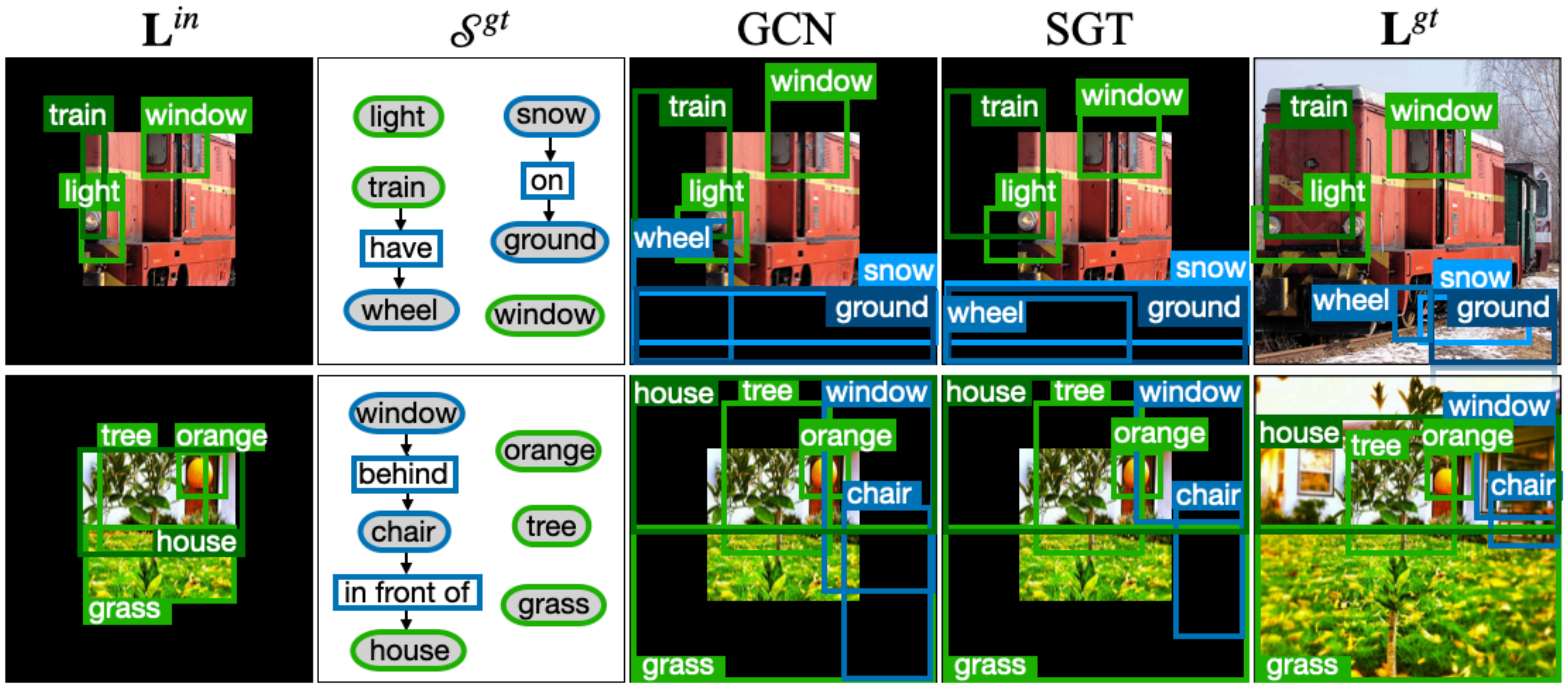
	VG-MSDN			
	Object		Relation	
	rAVG ↓	Hit@ 1 / 5 ↑	rAVG ↓	Hit@ 1 / 5 ↑
[1] Transformer	33.77	10.6 / 28.9	5.30	35.3 / 65.8
[2] LTNet	24.45	13.9 / 34.8	4.70	34.8 / 74.6
[3] GTwE	11.91	27.0 / 57.2	5.36	35.8 / 72.5
SGT	**8.38**	**39.7 / 68.9**	**3.43**	**55.3 / 84.3**
	COCO-stuff			
	Object		Relation	
	rAVG ↓	Hit@ 1 / 5 ↑	rAVG ↓	Hit@ 1 / 3 ↑
Transformer	22.35	14.7 / 37.8	2.37	29.4 / 78.5
LTNet	17.22	20.1 / 45.8	2.36	29.1 / 78.4
GTwE	11.81	28.4 / 57.2	2.89	20.4 / 63.3
SGT	**11.03**	**29.6 / 59.0**	**2.19**	**45.5 / 82.2**



[1] Vaswani et al. Attention is All you Need. NIPS, 2017.
[2] Yang et al. LayoutTransformer: Scene Layout Generation With Conceptual and Spatial Diversity. CVPR, 2021.
[3] Dwivedi et al. A Generalization of Transformer Networks to Graphs. AAAIW, 2021.

Experiments: G2L

	VG-MSDN	COCO-stuff
	mIoU	mIoU
[1] Transformer	5.1 / 71.2 / 51.9	10.4 / 75.7 / 61.2
[2] GCN	11.4 / 70.6 / 50.0	21.1 / 72.3 / 60.8
[3] GTwE	12.3 / 79.9 / 62.1	21.3 / 73.2 / 64.8
SGT	**14.5 / 81.1 / 62.4**	**28.2 / 85.1 / 74.9**



[1] Vaswani et al. Attention is All you Need. NIPS, 2017.
[2] Kipf et al. Semi-Supervised Classification with Graph Convolutional Networks. ICLR, 2016.
[3] Dwivedi et al. A Generalization of Transformer Networks to Graphs. AAAIW, 2021.

Experiments

- The accuracy of our semantic guidance extrapolation
- The quality of our outpainted images

Comparison

Input Boundless [1] AttSpade [2] Ours GT

[1] Teterwak et al. Boundless: Generative Adversarial Networks for Image Extension. ICCV, 2019.
[2] Herzig et al. Learning Canonical Representations for Scene Graph to Image Generation. ECCV, 2020

Comparison

[1] Teterwak et al. Boundless: Generative Adversarial Networks for Image Extension. ICCV, 2019.
[2] Herzig et al. Learning Canonical Representations for Scene Graph to Image Generation. ECCV, 2020

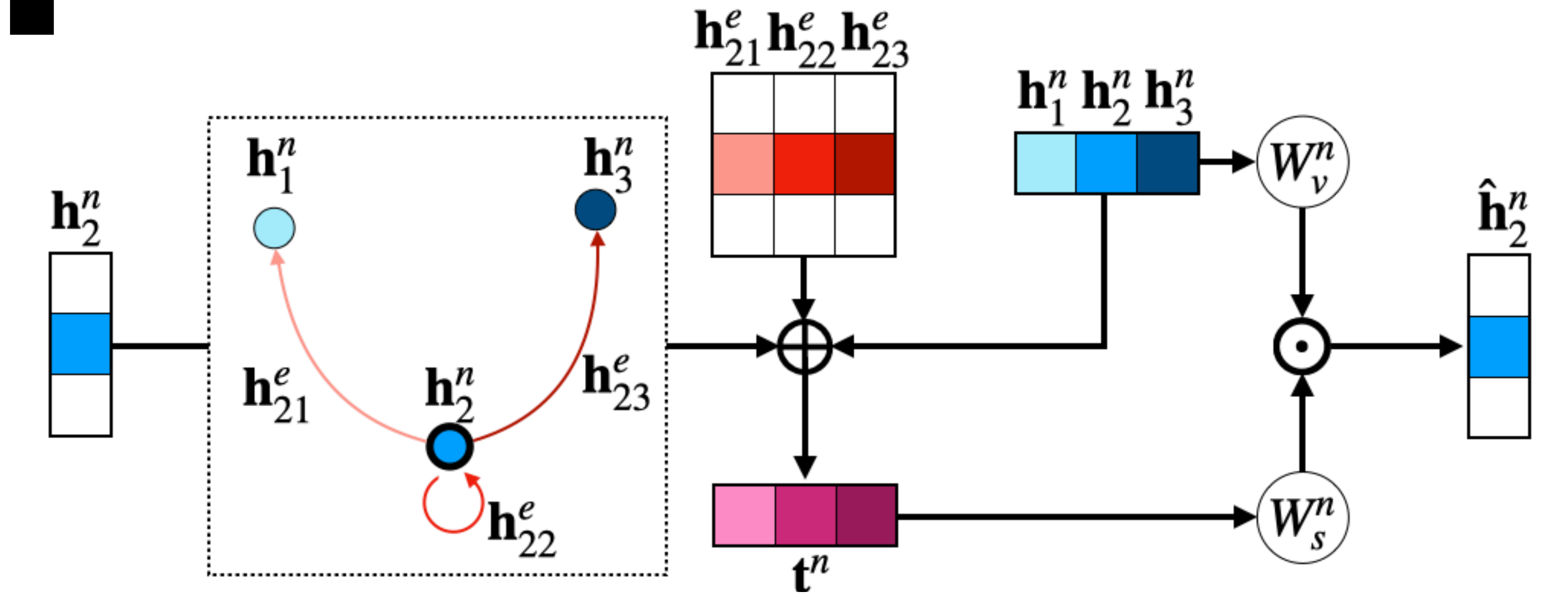
Semantic-guided Image Outpainting

Semantic-guided Image Outpainting

級一
• 內文層級二
級三
層級四
• 文層級五

Conclusion

- propose a novel **Scene Graph Transformer (SGT)** uniquely performs **attention** at both **node and edge levels** for modeling input structural information



- decompose the task into the stages **SGE**, **G2L**, and **L2I** leverage the information observed from the nodes and edges in the partial input scene graph, inferring plausible object co-occurrences, and thus producing the final image output