Complex Scene Image Editing by Scene Graph Comprehension

Zhongping Zhang, Huiwen He, Bryan A. Plummer, Zhenyu Liao, Huayan Wang

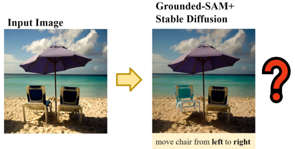
BOSTON UNIVERSITY · Kwai

BMVC 2023

Motivation

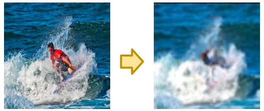
Challenges in Text-to-image editing:

- Accurately localizing and moving a target object in complex scenes can be challenging due to inherent ambiguities of text.



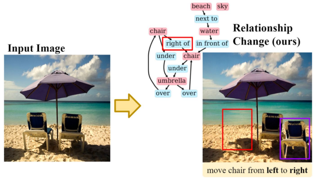
Which chair is the target object?

- GAN-based (one-stage) editing methods (*e.g.*, SIMSG) only support low resolution outputs and struggle to preserve the irrelevant attributes and details of the input image.



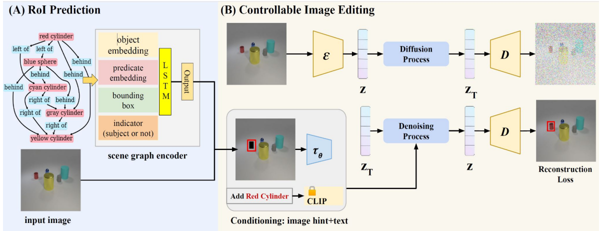
Our Two-stage Solution: SGC-Net

- First stage: localize and predict the desired position for target objects by **scene graph comprehension**;
- Second stage: achieve different editing operations with a **region-based image editing method.**



Use scene graph to identify the target object;
Use region-based diffusion model to perform editing operations.

Method



Our model mainly consists of two modules:

(A) **Region-of-Interest (RoI) prediction**: a RNN-based method to predict the desired regions for the target object with scene graph information.

Given modified triplets $\mathbf{y} = \{y_1, \ldots, y_T\}$ in the input scene graph, the encoded features can be obtained as:

$$m_t = concat\{V_s, V_o, V_p, b_s, I\}; \quad h_t = LSTM(m_t, h_{t-1})$$

- V_S, V_O, V_P: subject, object, and predicate embeddings, respectively;
- b_s: denotes the position of reference object;
- I : an indicator to indicate whether the target object is subject or object.

(B) **Region-based image editing:** a region-based image editing approach built on Stable Diffusion to achieve different image editing operations.

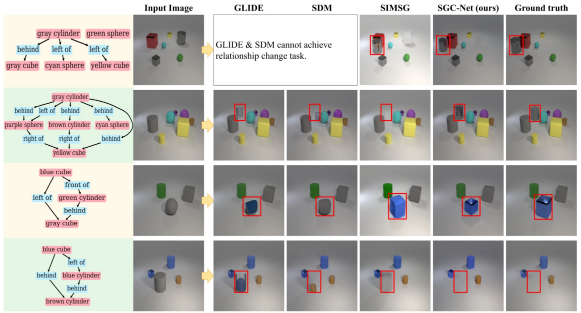
Quantitative Results (User Study)

User judgments on the correctness of an image manipulation on Visual Genome. Empty values indicate the approach is not capable of this task.

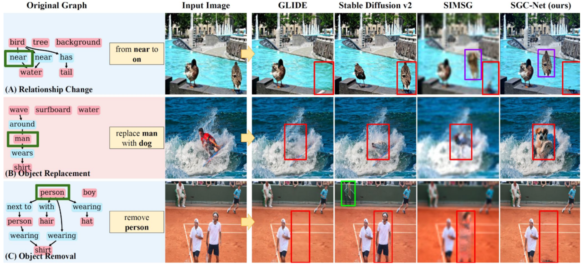
	Object Removal	Object Replacement	Relationship Change
GLIDE	23.2%	28.8%	-
Stable Diffusion 2	21.1%	32.2%	-
SIMSG	22.2%	24.4%	15.6%
SGC-Net (ours)	50.0%	40.0%	48.9%

Qualitative Results

- Examples on CLEVR



- Examples on Visual Genome



Takeaway

- Performing text-to-image editing **using scene graphs** can reduce manual effort and alleviate the issues caused by the ambiguity of text.
- A two stage model, performing **RoI prediction** and **region-based image editing** separately, can effectively perform various editing tasks in complex scenes and generate high-resolution (512x512) images.