Text-to-Image Synthesis based on Object-Guided Joint-Decoding Transformer

Fuxiang Wu[1,2], Liu Liu[3], Fusheng Hao[1,2], Fengxiang He[4], Jun Cheng[1,2*]
[1] Guangdong Provincial Key Laboratory of Robotics and Intelligent System,
Shenzhen Institute of Advanced Technology, CAS, China.
[2] The Chinese University of Hong Kong, Hong Kong, China.
[3] The University of Sydney, Australia.
[4] JD.com, Beijing, China.
{fx.wu1, fs.hao, jun.cheng}@siat.ac.cn, liu.liu1@sydney.edu.au, hefengxiang@jd.com.

1. Additional Experiments

We report more experiments about the controllable generation of our joint transformer by gradually including more objects in Figure 1 and Figure 2. In addition, we include more generated examples in Figure 3 and report more examples by using disturbing object labels in Figure 4 and by varying the bounding box of the largest object in Figure 5.

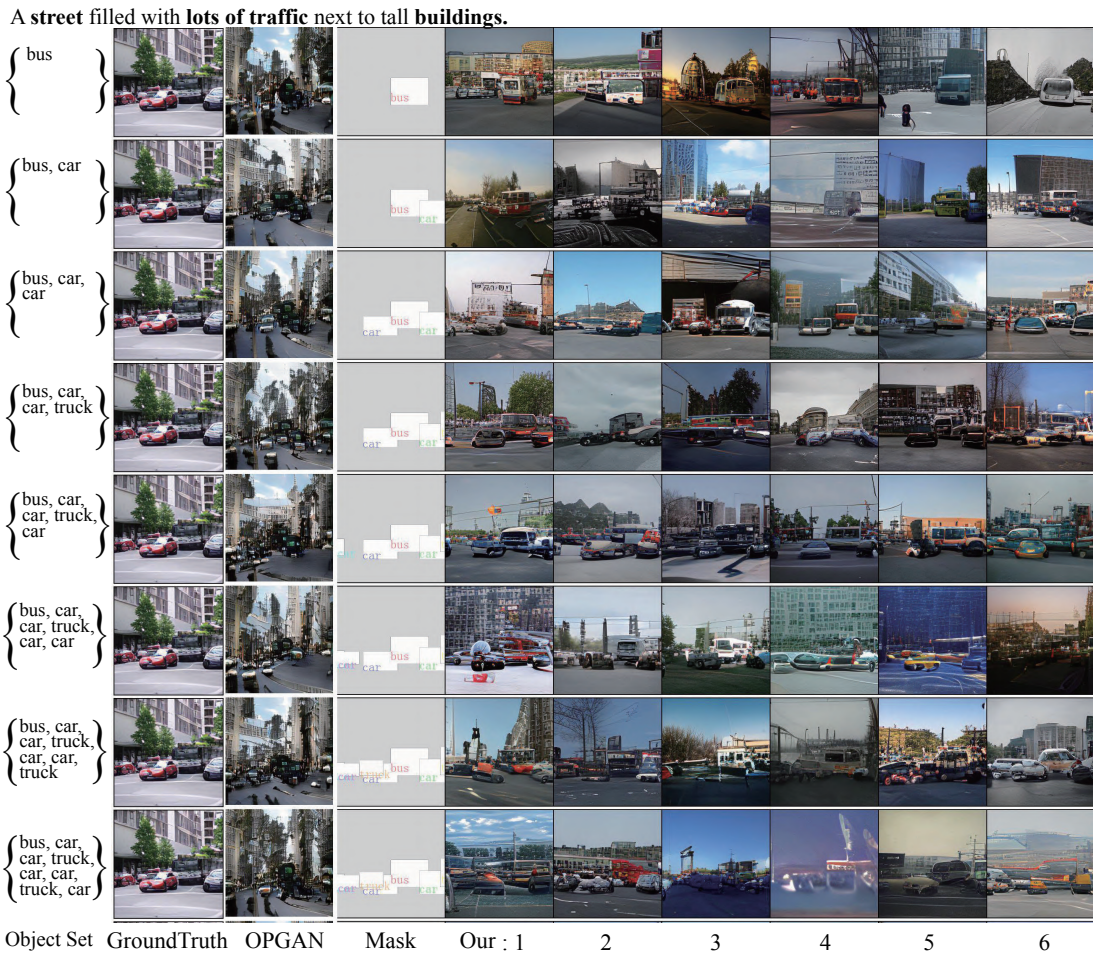

Figure 1. Generated examples with the variant objects depicted in the left brace: the caption is above the corresponding image and the prominent characteristics are marked as bold; "Mask" depicts the layout under the variant objects; "Our:1-6" are the synthesized samples.

Figure 2. Generated examples with the variant objects depicted in the left brace: the textual description is above the corresponding image and the prominent textual characteristics are marked as bold; "Mask" depicts the layout under the variant objects; "Our:1-6" are the synthesized samples.

Figure 3. Generated examples: the textual description is above the corresponding image and the prominent textual characteristics are marked as bold.

Figure 4. Generated examples with disturbing object labels: the textual description is above the corresponding image and the prominent textual characteristics are marked as bold; "Mask" depicts the layout under the disturbing object labels; "Our:1-6" are the synthesized samples.

A **stop sign** surrounded by overgrown bushes by a **railroad track.**

A **train** engine and **car** in a park display.

Figure 5. Generated examples with the largest object based on varied bounding boxes: the textual description is above the corresponding image and the prominent textual characteristics are marked as bold; "Mask" depicts the layout under the varied bounding boxes; "Our:1-6" are the synthesized samples.