A Additional Experiment Results

A.1 Qualitative Results.

We present additional qualitative results in Figure 1, 2, and 3 to supplement the main paper. The results demonstrate the effectiveness of SGC-Net in performing various editing tasks based on modified scene graphs. For example, in Figure 1, we see that SGC-Net produces plausible regions for the target object when the semantic relationships have been changed, such as “mirror – on – table”, “man – next to – wave”. The observation is consistent with our conclusion in the main paper.

A.2 Ablation Experiments on CLEVR.

Table 1 ablates our two modules. We find a significant gain in scene graph comprehension compared to text-only RoI prediction (71.50→79.48 on SSIM). In addition, our region-based editing module also boosts SDM (74.94→79.48), validating the effectiveness of our proposed modules.

<table>
<thead>
<tr>
<th>Method</th>
<th>MAE(RoI)</th>
<th>SSIM(RoI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SGC-Net(TEXT)</td>
<td>27.28</td>
<td>71.50</td>
</tr>
<tr>
<td>SGC-Net(SDM)</td>
<td>21.72</td>
<td>74.94</td>
</tr>
<tr>
<td>SGC-Net</td>
<td><strong>18.86</strong></td>
<td><strong>79.48</strong></td>
</tr>
</tbody>
</table>

Table 1: Ablation study on CLEVR. “TEXT” denotes text-only RoI prediction. “SDM” denotes Stable Diffusion [28].

© 2023. The copyright of this document resides with its authors. It may be distributed unchanged freely in print or electronic forms.
Figure 1: **Semantic Relationship Change.** Additional results of SGC-Net on the Visual Genome dataset. The modified nodes in scene graphs are outlined by red bounding boxes. We set the image resolution to $512 \times 512$ and simplify the scene graphs for better visualization. See Section A for discussion.
Figure 2: **Object Replacement.** Additional results of SGC-Net on the Visual Genome dataset. See Section A for discussion.
Figure 3: Object Removal. Additional results of SGC-Net on the Visual Genome dataset. See Section A for discussion.
In our user study, the annotators were shown an input image, a target text, and four edited images generated by different methods. The annotators were asked to choose which images accurately align with the target text. We provide a sample screenshot in Figure 4.

Figure 4: User study screenshot. A sample screenshot illustrating one of the questions presented to participants in our user study. See Section B for discussion.