A. ReCo with LAION data

In the main paper, we focus on the ReCo model trained on COCO (ReCo_{COCO}) to standardize the evaluation process. In this section, we present ReCo_{LAION} that conducts the same ReCo fine-tuning on a small subset of the LAION dataset [5] used by the pre-trained SD model [4]. Figure A shows selected ReCo_{LAION}-generated image samples.

Training setup. Instead of using the 414K image-text pairs (83K images) from the COCO 2014 training set, we randomly sample 100K images from the LAION-Aesthetics dataset¹. We take the Detic object detector [9] to generate

¹We use the first 100K samples with an aesthetics score of 6 or higher following the index in https://huggingface.co/datasets/ChristophSchuhmann/improved_aesthetics_6plus.
the object region predictions. We use a confidence threshold of 0.5 and filter out small boxes with a size smaller than 0.03 × W × H. Following the setting for ReCoCOCO, we feed all cropped regions to the pre-trained GIT captioning model [6] for regional descriptions. We fine-tune ReCo for 10,000 steps with the same training and inference settings introduced in the main paper.

Qualitative results. Figure B shows qualitative results on LVIS [2]. Both ReCoCOCO and ReCoLAION show strong region-controlled T2I generation capabilities. Compared with ReCoCOCO, ReCoLAION-generated images have better image aesthetic scores, thanks to the high-aesthetic fine-tuning data from LAION [5].

Figure C shows qualitative results on LAION-Aesthetics. We run T2I inference on 3K samples indexed after the first 100K samples used for ReCo fine-tuning.
ReCoLAION can preserve the pre-trained SD’s capabilities of understanding celebrities, art styles, and open-vocabulary descriptions, and meanwhile extend SD with the appealing new ability of region-controlled T2I generation.

Quantitative results. Table A compares ReCoLAION with ReCoCOCO on LVIS [2]. The “COCO Image” column indicates if the COCO image style is seen during ReCo fine-tuning. Automatic metrics show that ReCoCOCO achieves better region control accuracy and image FID. For region control, COCO ground-truth boxes provide a cleaner region specification than Detic-predicted boxes, thus benefiting the controlling accuracy. For the FID evaluation, ReCoCOCO has seen COCO images during ReCo training, leading to better FID scores. Qualitatively, ReCoLAION-generated images show comparable, if not better visual qualities than ReCoCOCO. Overall, both ReCo model variants significantly

A boat below a traffic light with a park in the background. <572> <576> <686> <314> a traffic light with the green light on. <298> <660> <730> <904> a white boat on the lake.

A zoomed out view of a man riding a horse through rural country side. <590> <699> <809> <854> brown horse. <672> <630> <721> <753> a man in blue shirt.

**Figure D.** Averaged ReCoCOCO cross-attention maps between visual latent and text embedding (on both text and position tokens).

<table>
<thead>
<tr>
<th>Method</th>
<th>COCO Image</th>
<th>Object Acc. (↑)</th>
<th>SceneFID (↓)</th>
<th>FID (↓)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real Images</td>
<td>-</td>
<td>42.00</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SD V1.4</td>
<td>×</td>
<td>7.88</td>
<td>40.62</td>
<td>23.74</td>
</tr>
<tr>
<td>ReCoCOCO</td>
<td>✓</td>
<td>23.42</td>
<td>10.08</td>
<td>17.73</td>
</tr>
<tr>
<td>ReCoLAION</td>
<td>×</td>
<td>19.38</td>
<td>19.48</td>
<td>21.99</td>
</tr>
</tbody>
</table>

**Table A.** Evaluations on the images generated with the 4,809 LVIS validation samples [2] from COCO val2017. The object classification is conducted over the 1,203 LVIS classes.

outperform the original SD model in both region control accuracy and image generation quality.

**B. Position Token Cross-Attention**

To help interpret how the introduced position tokens operate, Figure D visualizes the cross-attention maps between the visual latent $z$ and token embedding $\tau_y(y(P,T))$. We show the averaged attention maps across all diffusion steps and U-Net blocks. Similar to the cross-attention patterns observed in Pix2seq [1], we empirically observe that the four position tokens for each region help the model to progressively localize the specified area by attending to the corner or edge positions of the box region. These position tokens help text tokens to localize and focus on the detailed...
Figure E. ReCoOFA based on the auto-regressive T2I model OFA\textsubscript{large} [7].

Figure F. A case study of “inconsistent” image and region-level descriptions.

<table>
<thead>
<tr>
<th>Method</th>
<th>AP</th>
<th>AP(_{50})</th>
<th>Object Acc.</th>
<th>SceneFID</th>
<th>FID</th>
</tr>
</thead>
<tbody>
<tr>
<td>ReCo</td>
<td>32.0</td>
<td>52.4</td>
<td>62.42</td>
<td>6.51</td>
<td>7.36</td>
</tr>
<tr>
<td>ReCo\textsubscript{Position Word}</td>
<td>2.3</td>
<td>7.5</td>
<td>42.02</td>
<td>15.54</td>
<td>8.82</td>
</tr>
<tr>
<td>ReCo\textsubscript{Relation Word}</td>
<td>1.5</td>
<td>4.8</td>
<td>43.99</td>
<td>13.98</td>
<td>9.50</td>
</tr>
</tbody>
</table>

Table B. Extending Table 1 with additional ReCo model variants.

“Inconsistent” image and region-level descriptions. Supporting both image and region-level descriptions may raise a natural question: What if the image-level description is inconsistent with the region-level descriptions? For example, the image description might mention a dog, but none of the region descriptions refer to the dog. Since texts typically provide only partial descriptions of images or image patches, conflicts between image and region text descriptions may be rare in practice. As shown in Figure F, the model accommodates both image and region-level descriptions by properly drawing the dog outside of the box. We note that the model might still get confused with carefully engineered challenges, such as an image text stating “two dogs” paired with three “dog” regions. We leave those edge cases for future studies.

ReCo\textsubscript{Relation Word} with relationship words. In addition to the position text words used in ReCo\textsubscript{Position Word}, ReCo\textsubscript{Relation Word} further includes the eleven object spatial relationships and their textual descriptions defined in previous studies [3, 8]. Table B shows mixed results when compared with ReCo\textsubscript{Position Word}, while the performance remains lower than ReCo with position tokens. Therefore, we use ReCo\textsubscript{Position Word} as the reference model in the main paper and propose that position tokens could be inherently more concise and accurate for spatial controllability.

References

Figure G. Zooming in ReCoLaion-generated images shown in Figure B.
Figure H. Zooming in ReCoLAION-generated images shown in Figure C.


