ReCo: Region-Controlled Text-to-Image Generation

Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael Zeng, Lijuan Wang
Microsoft

{zhengyang,jianfw,zhe.gan,lindsey.li,keli,chewu,nanduan,zliu,ce.liu,nzeng,lijuanw}@microsoft.com

1. Introduction

Text-to-image (T2I) generation aims to generate faithful images based on an input text query that describes the image content. By scaling up the training data and model size, large T2I models [31, 34, 36, 45] have recently shown impressive performance in generating high-fidelity images, but with limited controllability, e.g., precisely specifying the content in a specific region with a free-form text description. In this paper, we propose an effective technique for such regional control in T2I generation. We augment T2I models' inputs with an extra set of position tokens, which represent quantized spatial coordinates. Each region is specified by four position tokens to represent the top-left and bottom-right corners, followed by an open-ended natural language regional description. Then, we fine-tune a pre-trained T2I model with such new input interface. Our model, dubbed as ReCo (Region-Controlled T2I), enables the region control for arbitrary objects described by open-ended regional texts rather than by object labels from a constrained category set. Empirically, ReCo achieves better image quality than the T2I model strengthened by positional words (FID: 8.82 → 7.36, SceneFID: 15.54 → 6.51 on COCO), together with objects being more accurately placed, amounting to a 20.40% region classification accuracy improvement on COCO. Furthermore, we demonstrate that ReCo can better control the object count, spatial relationship, and region attributes such as color/size, with the free-form regional description. Human evaluation on PaintSkill shows that ReCo is +19.28% and +17.21% more accurate in generating images with correct object count and spatial relationship than the T2I model. Code is available at https://github.com/microsoft/ReCo.

Abstract

Recently, large-scale text-to-image (T2I) models have shown impressive performance in generating high-fidelity images, but with limited controllability, e.g., precisely specifying the content in a specific region with a free-form text description. In this paper, we propose an effective technique for such regional control in T2I generation. We augment T2I models' inputs with an extra set of position tokens, which represent the quantized spatial coordinates. Each region is specified by four position tokens to represent the top-left and bottom-right corners, followed by an open-ended natural language regional description. Then, we fine-tune a pre-trained T2I model with such new input interface. Our model, dubbed as ReCo (Region-Controlled T2I), enables the region control for arbitrary objects described by open-ended regional texts rather than by object labels from a constrained category set. Empirically, ReCo achieves better image quality than the T2I model strengthened by positional words (FID: 8.82 → 7.36, SceneFID: 15.54 → 6.51 on COCO), together with objects being more accurately placed, amounting to a 20.40% region classification accuracy improvement on COCO. Furthermore, we demonstrate that ReCo can better control the object count, spatial relationship, and region attributes such as color/size, with the free-form regional description. Human evaluation on PaintSkill shows that ReCo is +19.28% and +17.21% more accurate in generating images with correct object count and spatial relationship than the T2I model. Code is available at https://github.com/microsoft/ReCo.

1. Introduction

Text-to-image (T2I) generation aims to generate faithful images based on an input text query that describes the image content. By scaling up the training data and model size, large T2I models [31, 34, 36, 45] have recently shown
remarkable capabilities in generating high-fidelity images. However, the text-only query allows limited controllability, e.g., precisely specifying the content in a specific region. The naive way of using position-related text words, such as “top left” and “bottom right,” often results in ambiguous and verbose input queries, as shown in Figure 2 (a). Even worse, when the text query becomes long and complicated, or describes an unusual scene, T2I models [31,45] might overlook certain details and rather follow the visual or linguistic training prior. These two factors together make region control difficult. To get the desired image, users usually need to try a large number of paraphrased queries and pick an image that best fits the desired scene. The process known as “prompt engineering” is time-consuming and often fails to produce the desired image.

The desired region-controlled T2I generation is closely related to the layout-to-image generation [7,9,22,23,34,38,44,49]. As shown in Figure 2 (b), layout-to-image models take all object bounding boxes with labels from a close set of object vocabulary [24] as inputs. Despite showing promise in region control, they can hardly understand free-form text inputs, nor the region-level combination of open-ended text descriptions and spatial positions. The two input conditions of text and box provide complementary referring capabilities. Instead of separately modeling them as in text-to-image and layout-to-image generations, we study “region-controlled T2I generation” that seamlessly combines these two input conditions. As shown in Figure 2 (c), the new input interface allows users to provide open-ended descriptions for arbitrary image regions, such as precisely placing a “brown glazed chocolate donut” in a specific area.

To this end, we propose ReCo (Region-Controlled T2I) that extends pre-trained T2I models to understand spatial coordinate inputs. The core idea is to introduce an extra set of input position tokens to indicate the spatial positions. The image width/height is quantized uniformly into \( N_{\text{bin}} \) bins. Then, any float-valued coordinate can be approximated and tokenized by the nearest bin. With an extra embedding matrix \( E_p \), the position token can be mapped onto the same space as the text token. Instead of designing a text-only query with positional words “in the top red donut” as in Figure 2 (a), ReCo takes region-controlled text inputs \( \langle x_1, y_1, x_2, y_2 \rangle \) red donut,” where \( x, y \) are the position tokens followed by the corresponding free-form text description. We then fine-tune a pre-trained T2I model with \( E_p \) to generate the image from the extended input query. To best preserve the pre-trained T2I capability, ReCo training is designed to be similar to the T2I pre-training, i.e., introducing minimal extra model parameters \( E_p \), jointly encoding position and text tokens with the text encoder, and prefixing the image description before the extended regional descriptions in the input query.

Figure 1 visualizes ReCo’s use cases and capabilities. As shown in Figure 1 (a), ReCo could reliably follow the input spatial constraints and generate the most plausible images by automatically adjusting object statues, such as the view (front/side) and type (single-/double-deck) of the “bus.” Position tokens also allow the user to provide free-form regional descriptions, such as “an orange cat wearing a red hat” at a specific location. Furthermore, we empirically observe that position tokens are less likely to get overlooked or misunderstood than text words. As shown in Figure 1 (b), ReCo has better control over object count, spatial relationships, and size properties, especially when the query is long and complicated, or describes a scene that is less common in real life. In contrast, T2I models [34] may struggle with generating scenes with correct object counts (“ten”), relationships (“boat below traffic light”), relative sizes (“chair larger than airplane”), and camera views (“zoomed out”).

To evaluate the region control, we design a comprehensive experiment benchmark based on a pre-trained regional object classifier and an object detector. The object classifier is applied on the generated image regions, while the detector is applied on the whole image. A higher accuracy means a better alignment between the generated object layout and the region positions in user queries. On the COCO dataset [24], ReCo shows a better object classification accuracy (42.02% \( \rightarrow \) 62.42%) and detector averaged precision (2.3 \( \rightarrow \) 32.0), compared with the T2I model with carefully designed positional words. For image generation quality, ReCo improves the FID from 8.82 to 7.36, and Scene-FID from 15.54 to 6.51. Furthermore, human evaluations on PaintSkill [5] show +19.28% and +17.21% accuracy gain in more correctly generating the query-described object count and spatial relationships, indicating ReCo’s capability in helping T2I models to generate challenging scenes.

Our contributions are summarized as follows.
• We propose ReCo that extends pre-trained T2I models to understand coordinate inputs. Thanks to the introduced position tokens in the region-controlled input query, users can easily specify free-form regional descriptions in arbitrary image regions.
• We instantiate ReCo based on Stable Diffusion. Extensive experiments show that ReCo strictly follows the regional instructions from the input query, and also generates higher-fidelity images.
• We design a comprehensive evaluation benchmark to validate ReCo’s region-controlled T2I generation capability. ReCo significantly improves both the region control accuracy and the image generation quality over a wide range of datasets and designed prompts.

2. Related Work

Text-to-image generation. Text-to-image (T2I) generation aims to generate a high-fidelity image based on an open-ended image description. Early studies adopt conditional GANs [33, 42, 46–48] for T2I generation. Recent studies have made tremendous advances by scaling up both the data and model size, based on either auto-regressive [10, 45] or diffusion-based models [31, 34, 36]. We build our study on top of the successful large-scale pre-trained T2I models, and explore how to better control the T2I generation by extending a pre-trained T2I model to understand position tokens.

Layout-to-image generation. Layout-to-image studies aim to generate an image from a complete layout, i.e., all bounding boxes and the paired object labels. Early studies [9, 22, 23, 38, 49] adopt GAN-based approaches by properly injecting the encoded layout as the input condition. Recent studies successfully apply the layout query as the input condition to the auto-regressive framework [10, 44] and diffusion models [7, 34]. Our study is related to the layout-to-image generation as both directions require the model to understand coordinate inputs. The major difference is that our design synergetically combines text and box to help T2I generation. Therefore, ReCo can take open-ended regional descriptions and benefit from large-scale T2I pre-training.

Unifying open-ended text and localization conditions. Previous studies have explored unifying open-ended text descriptions with localization referring (box, mask, mouse trace) as the input generation condition. One modeling approach [8, 10, 14, 17, 20, 28, 33] is to separately encode the image description in T2I and the layout condition in layout-to-image, and trains a model to jointly condition on both input types. TRECS [19] takes mouse traces in the localized narratives dataset [29] to better ground open-ended text descriptions with a localized position. Other than taking layout as user-generated inputs, previous studies [16, 21] have also explored predicting layout from text to ease the T2I generation of complex scenes. Unlike the motivation of training another conditional generation model parallel to T2I and layout-to-image, we explore how to effectively extend pre-trained T2I models to understand region queries, leading to significantly better controllability and generation quality than training from scratch. In short, we position ReCo as an improvement for T2I by providing a more flexible input interface and alleviating controllability issues, e.g., being difficult to override data prior when generating unusual scenes, and overlooking words in complex queries.

3. ReCo Model

Region-Controlled T2I Generation (ReCo) extends T2I models with the ability to understand coordinate inputs. The core idea is to design a unified input token vocabulary containing both text words and position tokens to allow accurate and open-ended regional control. By seamlessly mixing text and position tokens in the input query, ReCo obtains the best from the two worlds of text-to-image and layout-to-image, i.e., the abilities of free-form description and precise position control. In this section, we present our ReCo implementation based on the open-sourced Stable Diffusion (SD) [34]. We start with the SD preliminaries in Section 3.1 and introduce the core ReCo design in Section 3.2.

3.1. Preliminaries

We take Stable Diffusion as an example to introduce the T2I model that ReCo is built upon. Stable Diffusion is developed upon the Latent Diffusion Model [34], and consists of an auto-encoder, a U-Net [35] for noise estimation, and a CLIP ViT-L/14 text encoder. For the auto-encoder, the encoder $E$ with a down-sampling factor of 8 encodes the image $x$ into a latent representation $z = E(x)$ that the diffusion process operates on, and the decoder $D$ reconstructs the image $\hat{x} = D(z)$ from the latent $z$. U-Net [35] is conditioned on denoising timestep $t$ and text condition $\tau_0(y(T))$, where $y(T)$ is the input text query with text tokens $T$ and $\tau_0$ is the CLIP ViT-L/14 text encoder [30] that projects a sequence of tokenized texts into the sequence embedding.

The core motivation of ReCo is to explore more effective and interaction-friendly conditioning signals $y$, while best preserving the pre-trained T2I capability. Specifically, ReCo extends text tokens with an extra vocabulary specialized for spatial coordinate referring, i.e., position tokens $P$, which can be seamlessly used together with text tokens $T$ in a single input query $y$. ReCo aims to show the benefit of synergetically combining text and position conditions for region-controlled T2I generation.

3.2. Region-Controlled T2I Generation

ReCo input sequence. The text input in T2I generation provides a natural way of specifying the generation condition. However, text words may be ambiguous and verbose in providing regional specifications. For a better input
query, ReCo introduces position tokens that can directly refer to a spatial position. Specifically, the position and size of each region can be represented by four floating numbers, i.e., top-left and bottom-right coordinates. By quantizing coordinates [3,41,43], we can represent the region by four discrete position tokens \( P_i \), \(<x_1>, <y_1>, <x_2>, <y_2>\), arranged as a sequence similar to a short natural language sentence. The left side of Figure 3 illustrates the ReCo input sequence design. Same as T2I, we start the input query with the image description to make the best use of large-scale T2I pre-training. The image description is followed by multiple region-controlled texts, i.e., the four position tokens and the corresponding open-ended regional description. The number of regional specifications is unlimited, allowing users to easily create complex scenes with more regions, or save time on composing input queries with fewer or even no regions. ReCo introduces position token embedding \( E_P \in \mathbb{R}^{N_{pos} \times D} \) alongside the pre-trained text word embedding, where \( N_{pos} \) is the number of the position tokens, and \( D \) is the token embedding dimension. The entire sequence is then processed jointly, and each token, either text or spatial, is projected into a \( D \)-dim token embedding \( e \). The pre-trained CLIP text encoder from Stable Diffusion takes the token embeddings in, and projects them as the sequence embedding that the diffusion model conditions on.

ReCo fine-tuning. ReCo extends the text-only query \( y(T) \) with text tokens \( T \) into ReCo input query \( y(P,T) \) that combines the text word \( T \) and position token \( P \). We fine-tune the Stable Diffusion with the same latent diffusion modeling objective [34], following the notations in Section 3.1: 

\[
L = \mathbb{E}_{E(x),y(P,T),x \sim N(0,1),t} \left[ \| e - e_\theta(z_t, \tau_\theta(y(P,T))) \|_2^2 \right],
\]

where \( e_\theta \) and \( \tau_\theta \) are the fine-tuned network modules. All model parameters except position token embedding \( E_P \) are initiated from the pre-trained Stable Diffusion model. Both the image description and several regional descriptions are required for ReCo model fine-tuning. For the training data, we run a state-of-the-art captioning model [40] on the cropped image regions (following the annotated bounding boxes) to get the regional descriptions. During fine-tuning, we resize the image with the short edge to 512 and randomly crop a square region as the input image \( x \). We will release the generated data and fine-tuned model for reproduction.

We empirically observe that ReCo can well understand the introduced position tokens and precisely place objects at arbitrary specified regions. Furthermore, we find that position tokens can also help ReCo better model long input sequences that contain multiple detailed attribute descriptions, leading to fewer detailed descriptions being neglected or incorrectly generated than the text-only query. By introducing position tokens with a minimal change to the pre-trained T2I model, ReCo obtains the desired region controllability while best preserving the appealing T2I capability.

4. Experiments

4.1. Experiment Settings

ReCo takes region-controlled inputs specified by the users. However, gathering sufficient real user queries paired with images for quantitative evaluations may be challenging. Therefore, in addition to the arbitrarily-shaped boxes from PaintSkill [5] and manually designed challenging queries in Figure 7, we also include in-domain boxes from COCO [4] and LVIS [11] to construct evaluation queries.

Datasets. We quantitatively evaluate ReCo on COCO [4,24], PaintSkill [5], and LVIS [11]. For input queries, we take image descriptions and boxes from the datasets [5,11,24] and generate regional descriptions with GIT [40]. For COCO [4,24], we follow the established T2I setting [32,41,45] that reports the results on a subset of 30,000 captions sampled from the COCO 2014 val set. We fine-tune stable diffusion with image-text pairs from the COCO 2014
Table 1. Region control accuracy and image generation quality evaluations on the COCO (2014) 30k validation subset [24, 32, 41, 45].

<table>
<thead>
<tr>
<th>Method</th>
<th>Image Descr.</th>
<th>Region Descr.</th>
<th>Region Position</th>
<th>Region Control Metrics (†)</th>
<th>Image Quality Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>AP</td>
<td>AP50</td>
</tr>
<tr>
<td>Real Images</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>36.8</td>
<td>56.1</td>
</tr>
<tr>
<td>SD V1.4 Zero-shot</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>0.7</td>
<td>2.0</td>
</tr>
<tr>
<td>SD COCO Fine-tune:</td>
<td></td>
<td></td>
<td></td>
<td>0.6</td>
<td>1.8</td>
</tr>
<tr>
<td>ReCoImage Descr.</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>0.9</td>
<td>2.6</td>
</tr>
<tr>
<td>ReCoRegion Descr.</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>1.0</td>
<td>2.9</td>
</tr>
<tr>
<td>ReCoRegion Position Word</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>2.3</td>
<td>7.5</td>
</tr>
<tr>
<td>ReCo</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>32.0</td>
<td>52.4</td>
</tr>
</tbody>
</table>

Figure 4. Qualitative results on COCO [24]. ReCo’s extra regional control (shown in the dark blue color) can improve T2I generation on (a) object counting and relationships, (b) images with unique camera views, and (c) images with detailed regional attribute descriptions.

train set. PaintSkill [5] evaluates models’ capabilities on following arbitrarily-shaped boxes and generating images with the correct object type/count/relationship. We conduct the T2I inference with val set prompts, which contain 1,050/2,520/3,528 queries for object recognition, counting, and spatial relationship skills, respectively. LVIS [11] tests if the model understands open-vocabulary regional descriptions, with the object categories unseen in COCO fine-tuning. We report the results on the 4,809 LVIS val images [11] from the COCO 2017 val set [18, 49]. We do not fine-tune ReCo when experimenting on PaintSkill and LVIS to test its generalization capability with out-of-domain data.

Evaluation metrics. We evaluate ReCo with metrics on region control accuracy and image generation quality. For region control accuracy, we use Object Classification Accuracy [49] and DETR detector Average Precision (AP) [2]. Object accuracy trains a classifier with queried image crops to classify cropped regions on generated images. The generation model should generate objects in specified regions to obtain a high classification accuracy. DETR detector AP detects objects on generated images and compares the results with input object queries. Thus, higher accuracy and AP can indicate a better layout alignment. For image generation quality, we use the Fréchet Inception Distance (FID) [13] to evaluate the image quality. We take SceneFID [39] as an indicator for region-level visual quality, which computes FID on the regions cropped based on input object boxes. We compute FID and SceneFID with the Clean-FID repo [27] against center-cropped COCO images. We further conduct human evaluations on PaintSkill, due to the lack of GT images and effective automatic evaluation metrics.

Implementation details. We fine-tune ReCo from the Stable Diffusion v1.4 checkpoint. We introduce $N = 1000$ position tokens and increase the max length of the text encoder to 616. The batch size is 2048. We use AdamW optimizer [26] with a constant learning rate of $1e^{-4}$ to train the model for 20,000 steps, equivalent to around 100 epochs on COCO 2014 train set. The inference is conducted with 50 PLMS steps [25]. We select a classifier-free guidance scale [15] that gives the best region control performance, i.e., 4.0 for ReCo and 7.5 for original Stable Diffusion, detailed in Section 4.3. We do not use CLIP image re-ranking.

4.2. Region-Controlled T2I Generation Results

COCO. Table 1 reports the region-controlled T2I generation results on COCO. The first row “real images” provides an oracle reference number on applicable metrics. The top part of the table shows the results obtained with the pre-trained Stable Diffusion (SD) model without fine-tuning on COCO, i.e., the zero-shot setting. As shown in the left three columns, we experiment with adding “region description” and “region position” information to the input query in addition to “image description.” Since T2I models can not understand coordinates, we carefully design positional text de-
Figure 5. Qualitative results on PaintSkill [5]. ReCo’s extra regional control (shown in the dark blue color) can help T2I models more reliably generate scenes with exact object counts and unusual object relationships/relative sizes.

Table 2. Evaluations on the images generated with PaintSkill [5] prompts. We evaluate skill correctness with human judges, and object classification accuracy with the COCO-trained classifier.

(a) ReCo can more reliably generate images that involve counting or complex object relationships, e.g., “five birds” and “sitting on a bench.” (b) ReCo can more easily generate images with unique camera views by controlling the relative position and size of object boxes, e.g., “a top-down view of a cat” that T2I models struggle with. (c) Separating detailed regional descriptions with position tokens also helps ReCo better understand long queries and reduce attribute leakage, e.g., the color of the clock and person’s shirt.

PaintSkill. Table 2 shows the skill correctness and region control accuracy evaluations on PaintSkill [5]. Skill correctness [5] evaluates if the generated images contain the query-described object type/count/relationship, i.e., “object,” “count,” and “spatial” subsets. We use human judges to obtain the skill correctness accuracy. For region control, we use object classification accuracy to evaluate if the model follows those arbitrarily shaped and located object queries. We reuse the COCO region classifier introduced in Table 1.

Based on the human evaluation for “skill correctness,” 87.38% and 82.08% of ReCo-generated images have the correct object count and spatial relationship (“count” and “spatial”), which is +19.28% and +17.21% more accurate than ReCoPositionWord, and +26.98% and +32.97% higher than the T2I model with image description only. The skill correctness improvements suggest that region-control text inputs could be an effective interface to help T2I models more reliably generate user-specified scenes. The object accuracy evaluation makes the criteria more strict by requiring the model to follow the exact input region positions, in ad-
ReCo can understand open-vocabulary regional descriptions, including keywords such as "curtain," "ferris wheel," "sausage," and "salt shaker."

PaintSkill contains input queries with randomly assigned object types, locations, and shapes. Because of the minimal constraints, many queries describe challenging scenes that appear less frequently in real life. We observe that ReCo not only precisely follows position queries, but also fits objects and their surroundings naturally, indicating an understanding of object properties. In Figure 5 (a), the three buses with different aspect ratios each have their unique viewing angle and direction, such that the object "bus" fits tightly with the given region. More interestingly, the directions of each bus go nicely with the road, making the image look real to humans. Figure 5 (b) shows challenging cases that require drawing two less commonly co-occurred objects into the same image. ReCo correctly fits "bed" and "fire hydrant," "boat" and "bus" into the given region. More impressively, ReCo can create a scene that makes the generated image look plausible, e.g., "looking through a window with a bed indoors," with the commonsense knowledge that "bed" is usually indoor while "fire hydrant" is usually outdoor. The randomly assigned region categories can also lead to objects with unusual relative sizes, e.g., the bag that is larger than the airplane in Figure 5 (c). ReCo shows an understanding of image perspectives by placing smaller objects such as "backpack" and "dog" near the camera position.

**LVIS.** Table 3 reports the T2I generation results with out-of-vocabulary regional entities. We observe that ReCo can understand open-vocabulary regional descriptions, by transferring the open-vocab capability learned from large-scale T2I pre-training to regional descriptions. ReCo achieves the best SceneFID and object classification accuracy over the 1,203 LVIS classes of 10.08 and 23.42%. The results show that the ReCo position tokens can be used with open-vocabulary regional descriptions, despite being trained on COCO with 80 object types. Figure 6 shows examples of generating objects that are not annotated in COCO, e.g., "curtain" and "loveseat" in (a), "ferris wheel" and "clock tower" in (b), "sausage" and "tomato" in (c), "salts" in (d).

**Qualitative results.** We next qualitatively show ReCo’s other capabilities with manually designed input queries. Figure 1 (a) shows examples of arbitrary object manipulation and regional description control. As shown in the "bus" example, ReCo will automatically adjust the object viewing (from side to front) and type (from single- to double-deck) to reasonably fit the region constraint, indicating the knowledge about object "bus." ReCo can also understand the free-form regional text and generate "cats" in the specified region with different attributes, e.g., "wearing a red hat," "pink," "sleeping," etc. Figure 7 (a) shows an example of generating images with different object counts. ReCo’s region control provides a strong tool for generating the exact object count, optionally with extra regional texts describing each object. Figure 7 (b) shows how we can use the box size to control the camera view, e.g., the precise control of the exact zoom-in ratio. Figure 7 (c) presents additional examples of images with unusual object relationships.

### 4.3. Analysis

**Regional descriptions.** Alternative to the open-ended free-form texts, regional descriptions can be object indexes from a constrained category set, as the setup in layout-to-image generation [9, 22, 23, 38, 49]. Table 4 compares ReCo with ReCoODLabel on COCO [24] and LVIS [11]. The leftmost “accuracy” column on COCO shows the major advantage of ReCoODLabel, i.e., when fine-tuned and tested with the same regional object vocabulary, ReCoODLabel is +7.28% higher in region control accuracy, compared with ReCo. However, the closed-vocabulary OD labels bring two disadvantages. First, the position tokens in ReCoODLabel tend to only work with the seen vocabulary, i.e., the 80 COCO categories. When evaluated on other datasets such as LVIS or
open-world use cases, the region control performance drops significantly, as shown in the “accuracy” column on LVIS. Second, ReCo\textsubscript{OD Label} only works well with constrained object labels, which fail to provide detailed regional descriptions, such as attributes and object relationships. Therefore, ReCo\textsubscript{OD Label} helps less in generating high-fidelity images, with FID 1.72 and 5.33 worse than ReCo on COCO and LVIS. Given the aforementioned limitations, we use the open-ended free-form regional descriptions in ReCo.

Guidance scale and T2I SOTA comparison. Table 5 (a,b) examines how different classifier-free guidance scales [15] influence region control accuracy and image generation quality on the COCO 2014 val subset [24, 32, 41, 45]. We empirically observe that scale of 1.5 yields the best image quality, and a slightly larger scale of 4.0 provides the best region control performance. Table 5 (c) compares ReCo with the state-of-the-art T2I methods in the fine-tuned setting. We reduce the guidance scale from the 4.0 in Table 1 to 1.5 for a fair comparison. We do not use any image-text contrastive models for results re-ranking. ReCo achieves an FID of 5.18, compared with 6.98 when we fine-tune Stable Diffusion with COCO T2I data without regional description. ReCo also outperforms the real image retrieval baseline [45] and most prior studies [6, 10, 46, 50].

Limitations. Our method has several limitations. First, ReCo might generate lower-quality images when the input query becomes too challenging, e.g., the unusual giant “dog” in Figure 7 (c). Second, for evaluation purposes, we train ReCo on the COCO train set. Despite preserving the open-vocabulary capability shown on LVIS, the generated image style does bias towards COCO. This limitation can potentially be alleviated by conducting the same ReCo fine-tuning on a small subset of pre-training data [37] used by the same T2I model [34]. We show this ReCo variant in the supplementary material. Finally, ReCo builds upon large-scale pre-trained T2I models such as Stable Diffusion [34] and shares similar possible generation biases.

5. Conclusion

We have presented ReCo that extends a pre-trained T2I model for region-controlled T2I generation. Our introduced position token allows the precise specification of open-ended regional descriptions on arbitrary image regions, leading to an effective new interface of region-controlled text input. We show that ReCo can help T2I generation in challenging cases, e.g., when the input query is complicated with detailed regional attributes or describes an unusual scene. Experiments validate ReCo’s effectiveness on both region control accuracy and image generation quality.
References


