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

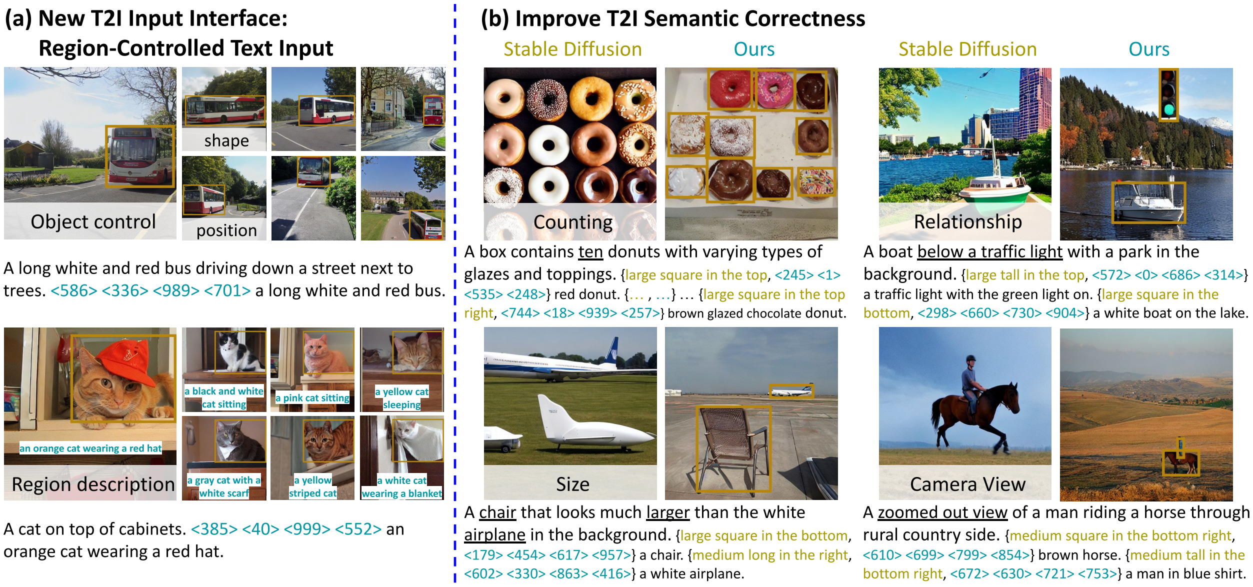

(a) New T2I Input Interface:
Region-Controlled Text Input
Object control
shape
position
A long white and red bus driving down a street next to trees. <586> <336> <989> <701> a long white and red bus.
an orange cat wearing a red hat
Region description
a black and white cat sitting
a pink cat sitting
a yellow cat sleeping
a gray cat with a white scarf
a yellow striped cat
a white cat wearing a blanket
A cat on top of cabinets. <385> <40> <999> <552> an orange cat wearing a red hat.
(b) Improve T2I Semantic Correctness
Stable Diffusion
Ours
Stable Diffusion
Ours
Counting
A box contains ten donuts with varying types of glazes and toppings. {large square in the top, <245> <1> <535> <248>} red donut. {… , …} … {large square in the top right, <744> <18> <939> <257>} brown glazed chocolate donut.
Relationship
A boat below a traffic light with a park in the background. {large tall in the top, <572> <0> <686> <314>} a traffic light with the green light on. {large square in the bottom, <298> <660> <730> <904>} a white boat on the lake.
Size
A chair that looks much larger than the white airplane in the background. {large square in the bottom, <179> <454> <617> <957>} a chair. {medium long in the right, <602> <330> <863> <416>} a white airplane.
Camera View
A zoomed out view of a man riding a horse through rural country side. {medium square in the bottom right, <610> <699> <799> <854>} brown horse. {medium tall in the bottom right, <672> <630> <721> <753>} a man in blue shirt.

Figure 1. **(a)** *ReCo* extends pre-trained text-to-image models (Stable Diffusion [34]) with an extra set of input position tokens (in dark blue color) that represent quantized spatial coordinates. Combining position and text tokens yields the region-controlled text input, which can specify an open-ended regional description precisely for any image region. **(b)** With the region-controlled text input, *ReCo* can better control the object count/relationship/size properties and improve the T2I semantic correctness. We empirically observe that position tokens are less likely to get overlooked than positional text words, especially when the input query is complicated or describes an unusual scene.

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: $8.82 \rightarrow 7.36$, SceneFID: $15.54 \rightarrow 6.51$ 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_{bins} 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 "$<x_1>$, $<y_1>$, $<x_2>$, $<y_2>$ 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 relationship, 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").

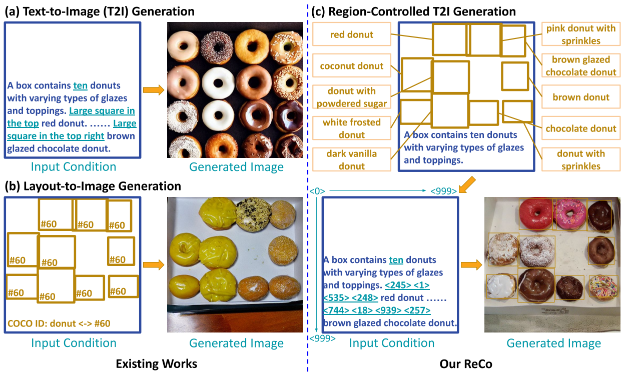

(a) Text-to-Image (T2I) Generation
A box contains ten donuts with varying types of glazes and toppings. Large square in the top red donut. …… Large square in the top right brown glazed chocolate donut.
Input Condition
Generated Image
(b) Layout-to-Image Generation
#60
#60
#60
#60
#60
#60
#60
#60
#60
#60
COCO ID: donut <-> #60
Input Condition
Generated Image
Existing Works
(c) Region-Controlled T2I Generation
red donut
coconut donut
donut with powdered sugar
white frosted donut
dark vanilla donut
pink donut with sprinkles
brown glazed chocolate donut
brown donut
chocolate donut
donut with sprinkles
A box contains ten donuts with varying types of glazes and toppings.
<0>
<999>
<999>
A box contains ten donuts with varying types of glazes and toppings. <245> <1> <535> <248> red donut …… <744> <18> <939> <257> brown glazed chocolate donut.
Input Condition
Generated Image
Our ReCo

Figure 2. (a) With positional words (*e.g.*, bottom/top/left/right and large/small/tall/long), the T2I model (Stable Diffusion [34]) does not manage to create objects with desired properties. **(b)** Layout-to-image generation [22, 34, 38, 49] takes all object boxes and labels as the input condition, but only works well with constrained object labels. **(c)** Our ReCo model synergetically combines the text and box referring, allowing users to specify an open-ended regional text description precisely at any image region.

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 relationship, 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 $\mathcal{E}$ with a down-sampling factor of 8 encodes the image x into a latent representation $z = \mathcal{E}(x)$ that the diffusion process operates on, and the decoder $\mathcal{D}$ reconstructs the image $\hat{x} = \mathcal{D}(z)$ from the latent z. U-Net [35] is conditioned on denoising timestep t and text condition $\tau_\theta(y(T))$, where $y(T)$ is the input text query with text tokens T and τ_θ 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

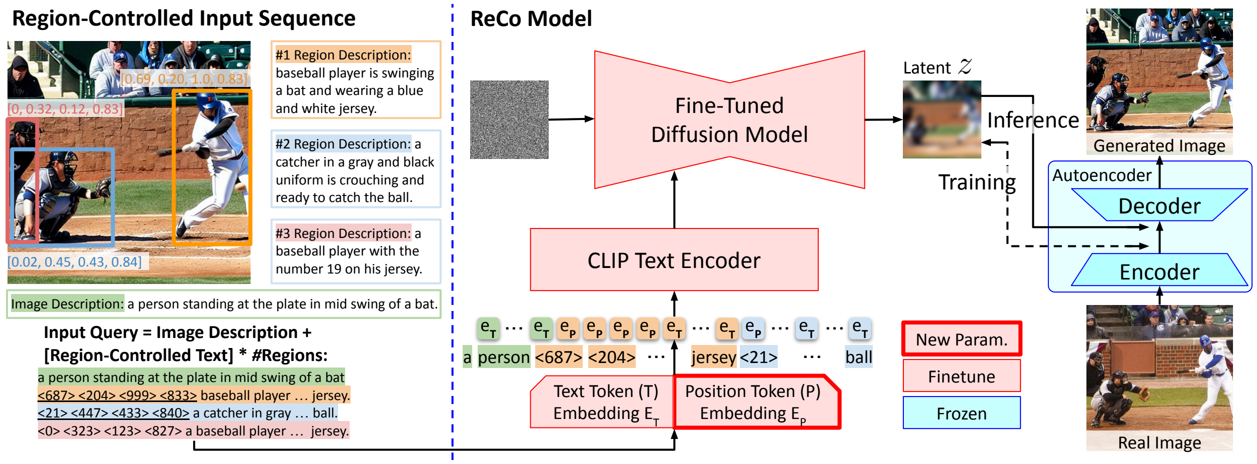

Region-Controlled Input Sequence
[0.69, 0.20, 1.0, 0.83]
[0, 0.32, 0.12, 0.83]
[0.02, 0.45, 0.43, 0.84]
#1 Region Description: baseball player is swinging a bat and wearing a blue and white jersey.
#2 Region Description: a catcher in a gray and black uniform is crouching and ready to catch the ball.
#3 Region Description: a baseball player with the number 19 on his jersey.
Image Description: a person standing at the plate in mid swing of a bat.
Input Query = Image Description + [Region-Controlled Text] * #Regions:
a person standing at the plate in mid swing of a bat
<687> <204> <999> <833> baseball player … jersey.
<21> <447> <433> <840> a catcher in gray … ball.
<0> <323> <123> <827> a baseball player … jersey.
ReCo Model
Fine-Tuned Diffusion Model
Latent z
Inference
Training
Generated Image
Autoencoder
Decoder
Encoder
Real Image
CLIP Text Encoder
e_T ··· e_T e_P e_P e_P e_P e_T ··· e_T e_P ··· e_T ··· e_T
a person <687> <204> ··· jersey <21> ··· ball
Text Token (T) Embedding E_T
Position Token (P) Embedding E_P
New Param.
Finetune
Frozen

Figure 3. ReCo extends Stable Diffusion [34] with position tokens P to support open-ended text description at both image- and region-level. We minimize the amount of introduced new model parameters (*i.e.*, position token embedding E_P) to best preserve the pre-trained T2I capability. The diffusion model and text encoder are fine-tuned together to support the extended position token inputs.

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, $<x_1>$, $<y_1>$, $<x_2>$, $<y_2>$, 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 $E_P \in \mathbb{R}^{N_{\text{bins}} \times D}$ alongside the pre-trained text word embedding, where N_{bins} 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:

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

where ϵ_θ and τ_θ 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-part 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

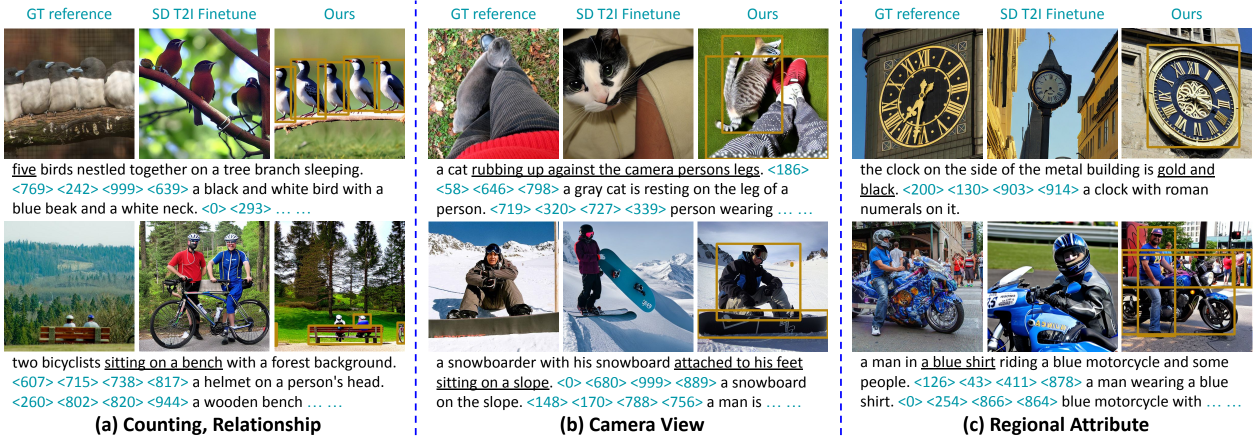
4.1. Experiment Settings

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

Method	Image Descr.	Region Descr.	Region Position	Region Control Metrics (↑) AP	AP$_{50}$	Object Acc.	Image Quality Metrics SceneFID (↓)	FID (↓)
Real Images	-	-	-	36.8	56.1	71.41	-	-
SD V1.4 Zero-shot	✓	-	-	0.7	2.0	26.75	35.80	13.40
	✓	✓	-	0.7	2.0	27.64	34.72	13.88
	✓	✓	Text	0.6	1.8	28.15	32.86	14.57
SD COCO Fine-tune:								
ReCo$_{\text{Image Descr.}}$	✓	-	-	0.9	2.6	29.12	27.78	10.44
ReCo$_{\text{Region Descr.}}$	✓	✓	-	1.0	2.9	32.32	24.88	9.11
ReCo$_{\text{Position Word}}$	✓	✓	Text	2.3	7.5	42.02	15.54	8.82
ReCo	✓	✓	✓	**32.0**	**52.4**	**62.42**	**6.51**	**7.36**

Table 1. Region control accuracy and image generation quality evaluations on the COCO (2014) 30k validation subset [24, 32, 41, 45].



five birds nestled together on a tree branch sleeping. <769> <242> <999> <639> a black and white bird with a blue beak and a white neck. <0> <293> … …

a cat rubbing up against the camera persons legs. <186> <58> <646> <798> a gray cat is resting on the leg of a person. <719> <320> <727> <339> person wearing … …

the clock on the side of the metal building is gold and black. <200> <130> <903> <914> a clock with roman numerals on it.

two bicyclists sitting on a bench with a forest background. <607> <715> <738> <817> a helmet on a person's head. <260> <802> <820> <944> a wooden bench … …

a snowboarder with his snowboard attached to his feet sitting on a slope. <0> <680> <999> <889> a snowboard on the slope. <148> <170> <788> <756> a man is … …

a man in a blue shirt riding a blue motorcycle and some people. <126> <43> <411> <878> a man wearing a blue shirt. <0> <254> <866> <864> blue motorcycle with … …

(a) Counting, Relationship **(b) Camera View** **(c) Regional Attribute**

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.

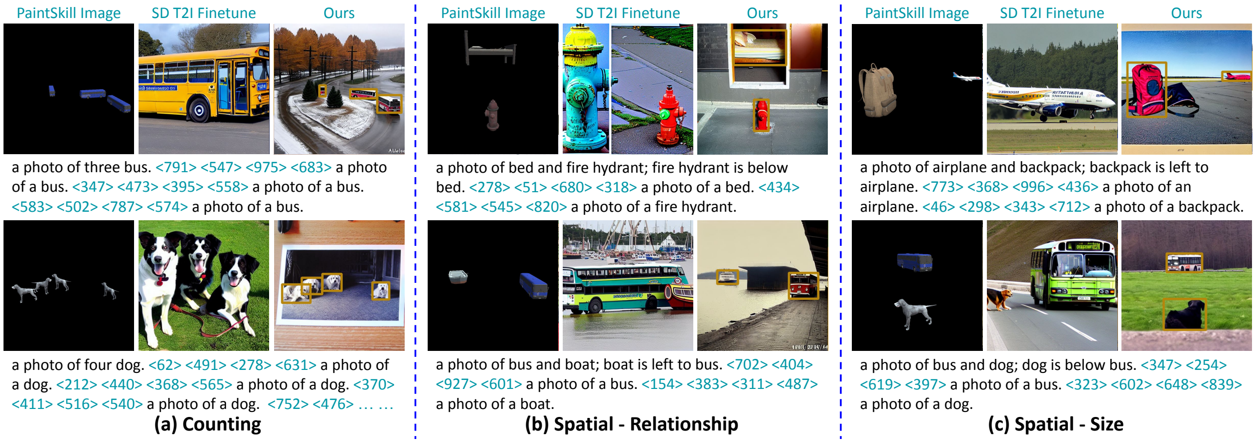
train set. PaintSkill [5] evaluates models' capabilities on following arbitrarily-shaped boxes and generating images with the correct object type/count/relationship. We conduct the T2I inference with val set prompts, which contain 1,050/2,520/3,528 queries for object recognition, counting, and spatial 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-


PaintSkill Image
SD T2I Finetune
Ours
PaintSkill Image
SD T2I Finetune
Ours
PaintSkill Image
SD T2I Finetune
Ours
a photo of three bus. <791> <547> <975> <683> a photo of a bus. <347> <473> <395> <558> a photo of a bus. <583> <502> <787> <574> a photo of a bus.
a photo of bed and fire hydrant; fire hydrant is below bed. <278> <51> <680> <318> a photo of a bed. <434> <581> <545> <820> a photo of a fire hydrant.
a photo of airplane and backpack; backpack is left to airplane. <773> <368> <996> <436> a photo of an airplane. <46> <298> <343> <712> a photo of a backpack.
a photo of four dog. <62> <491> <278> <631> a photo of a dog. <212> <440> <368> <565> a photo of a dog. <370> <411> <516> <540> a photo of a dog. <752> <476> … …
a photo of bus and boat; boat is left to bus. <702> <404> <927> <601> a photo of a bus. <154> <383> <311> <487> a photo of a boat.
a photo of bus and dog; dog is below bus. <347> <254> <619> <397> a photo of a bus. <323> <602> <648> <839> a photo of a dog.

(a) Counting

(b) Spatial - Relationship

(c) Spatial - Size

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.

scriptions, indicated by "text" in the "region position" column. Specifically, we describe a region with one of the three size words (*small, medium, large*), three possible region aspect ratios (*long, square, tall*), and nine possible locations (*top left, top, . . ., bottom right*). We note that the resulting ReCo$_{\text{Position Word}}$ serves as a strengthened T2I baseline for reference purposes, and it is unfair to directly compare it with ReCo that understands coordinates. The *bottom part* compares the main ReCo model with other variants finetuned with the corresponding input queries. The *middle three* rows report the results on region control accuracy. For AP and AP$_{50}$, we use a DETR ResNet-50 object detector trained on COCO [1, 2] to get the detection results on images generated based on the input texts and boxes from the COCO 2017 val5k set [24]. The "object accuracy" column reports the region classification accuracy [49]. The trained ResNet-101 region classifier [12] yields a 71.41% oracle 80-class accuracy on real images. The *right two* columns report the image generation quality metrics SceneFID a nd FID, which evaluate the region and image visual qualities.

One advantage of ReCo is its strong region control capability. As shown in the bottom row, ReCo achieves an AP of 32.0, which is close to the real image oracle of 36.8. Despite the careful engineering of positional text words, ReCo$_{\text{Position Word}}$ only achieves an AP of 2.3. Similarly, for object region classification, 62.42% of the cropped regions on ReCo-generated images can be correctly classified, compared with 42.02% of ReCo$_{\text{Position Word}}$. ReCo also improves the generated image quality, both at the region and image level. At the region level, ReCo achieves a SceneFID of 6.51, indicating strong capabilities in both generating high-fidelity objects and precisely placing them in the queried position. At the image level, ReCo improves the FID from 10.44 to 7.36 with the region-controlled text input that provides a localized and more detailed image description. We present additional FID comparisons to state-of-the-art conditional image generation methods in Table 5 (c).

We show representative qualitative results in Figure 4.

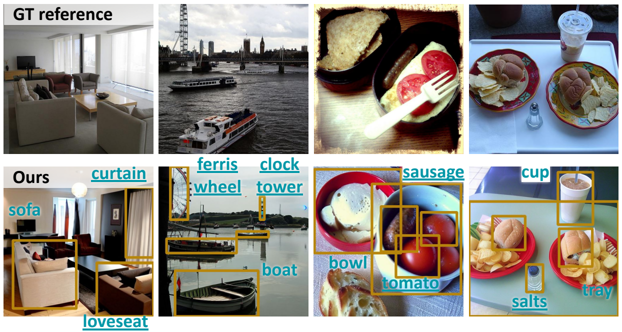
Method	Skill Correctness (↑)			Object Accuracy (↑)		
	Object	Count	Spatial	Object	Count	Spatial
SD V1.4 Zero-shot	97.11	59.28	48.20	35.97	22.32	13.06
ReCo$_{\text{Image Descr.}}$	98.23	60.40	49.11	38.92	25.79	15.17
ReCo$_{\text{Position Word}}$	93.33	68.10	64.87	50.72	25.35	22.82
ReCo	**98.51**	**87.38**	**82.08**	**82.30**	**63.40**	**67.30**

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.*, the "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 ReCo$_{\text{Position Word}}$, 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-



(a) a living room that has a bunch of different couches. <816> <140> <999> <607> white curtain is shown in a room. <622> <642> …

(b) there are two large boats that are in the water. <80> <0> <196> <354> a ferris wheel is in the sky with a black background. <676> …

(c) a lunch bowl with bread, sausage, tomato, and egg. <450> <267> <692> <567> a close up of a sausage and tomato. <550> <459> …

(d) a tray topped with two plates of food and a drink. <383> <653> <496> <836> a clear glass salt shaker. <158> <331> <375> <555> …

Figure 6. Qualitative results on LVIS [11]. ReCo can understand open-vocabulary regional descriptions, including keywords such as “curtain,” “ferris wheel,” “sausage,” and “salt shaker.”

Method	Object Acc. (↑)	SceneFID (↓)	FID (↓)
Real Images	42.00	-	-
SD V1.4 Zero-shot	7.88	40.62	23.74
ReCo$_{\text{Image Descr.}}$	9.82	28.95	20.87
ReCo$_{\text{Region Descr.}}$	11.08	28.15	17.96
ReCo$_{\text{Position Word}}$	16.60	20.27	17.80
ReCo	**23.42**	**10.08**	**17.73**

Table 3. Evaluations on the images generated with the 4,809 LVIS validation samples [11] from COCO val2017. The object classification is conducted over the 1,203 LVIS classes.

Method	COCO			LVIS		
	Acc.	SceneFID	FID	Acc.	SceneFID	FID
Real Images	74.41	-	-	42.00	-	-
ReCo$_{\text{OD Label}}$	**69.70**	8.07	9.08	22.79	13.98	23.06
ReCo	62.42	**6.51**	**7.36**	**23.42**	**10.08**	**17.73**

Table 4. Analyses on using open-ended texts (ReCo) *vs.* constrained object labels (ReCo$_{\text{OD Label}}$) as the regional description.

dition to skills. “ReCo” achieves a strong region control accuracy of 63.40% and 67.30% on count and skill subsets, surpassing “ReCo$_{\text{Position Word}}$” by $+38.05\%$ and $+44.48\%$.

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 ReCo$_{\text{OD Label}}$ on COCO [24] and LVIS [11]. The leftmost “accuracy” column on COCO shows the major advantage of ReCo$_{\text{OD Label}}$, *i.e.*, when fine-tuned and tested with the same regional object vocabulary, ReCo$_{\text{OD Label}}$ 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 ReCo$_{\text{OD Label}}$ tend to only work with the seen vocabulary, *i.e.*, the 80 COCO categories. When evaluated on other datasets such as LVIS or

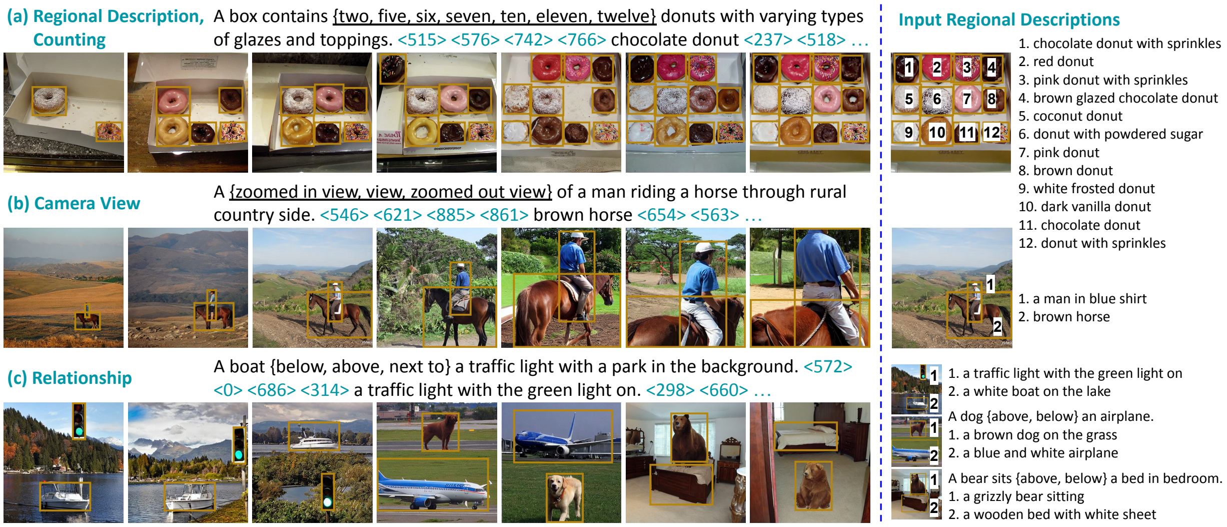
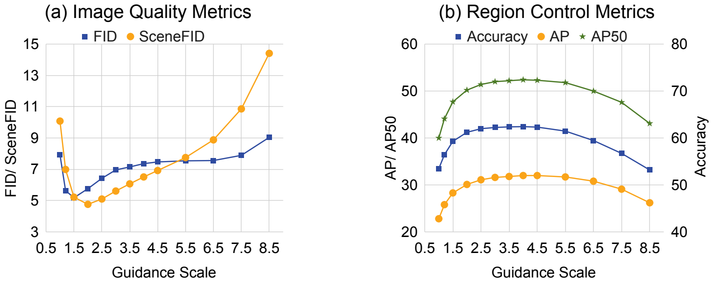

(a) Regional Description, Counting
A box contains {two, five, six, seven, ten, eleven, twelve} donuts with varying types of glazes and toppings. <515> <576> <742> <766> chocolate donut <237> <518> …
(b) Camera View
A {zoomed in view, view, zoomed out view} of a man riding a horse through rural country side. <546> <621> <885> <861> brown horse <654> <563> …
(c) Relationship
A boat {below, above, next to} a traffic light with a park in the background. <572> <0> <686> <314> a traffic light with the green light on. <298> <660> …
Input Regional Descriptions
1 2 3 4 5 6 7 8 9 10 11 12
1. chocolate donut with sprinkles
2. red donut
3. pink donut with sprinkles
4. brown glazed chocolate donut
5. coconut donut
6. donut with powdered sugar
7. pink donut
8. brown donut
9. white frosted donut
10. dark vanilla donut
11. chocolate donut
12. donut with sprinkles
1 2
1. a man in blue shirt
2. brown horse
1 2
1. a traffic light with the green light on
2. a white boat on the lake
A dog {above, below} an airplane.
1. a brown dog on the grass
2. a blue and white airplane
A bear sits {above, below} a bed in bedroom.
1. a grizzly bear sitting
2. a wooden bed with white sheet

Figure 7. Qualitative results of ReCo-generated images with manually designed challenging input queries.



(c) Method	FID (↓)
Random Train Images [10]	2.47
Retrieval Baseline [45]	6.82
XMC-GAN [46]	9.33
CogView2 [6]	17.7
LAFITE [50]	8.12
Make-A-Scene [10]	7.55
Parti [45]	3.22
ReCo$_{\text{Image Descr.}}$	6.98
ReCo$_{\text{Position Word}}$	5.98
ReCo	5.18

Table 5. (a,b) Analyses of different guidance scales' influences [15] on image quality and region control accuracy. **(c)** Comparison with previous T2I works on the COCO (2014) validation 30k subset [24, 32, 41, 45]) in the fine-tuned setting.

open-world use cases, the region control performance drops significantly, as shown in the "accuracy" column on LVIS. *Second*, ReCo$_{\text{OD Label}}$ only works well with constrained object labels, which fail to provide detailed regional descriptions, such as attributes and object relationships. Therefore, ReCo$_{\text{OD Label}}$ 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

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. Detr res50 checkpoint from the official detr repo. In *https://dl.fbaipublicfiles.com/detr/detr-r50-e632da11.pth*, 2020. 6

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 5, 6

[3] Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. In *ICLR*, 2022. 4

[4] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 4

[5] Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers. *arXiv preprint arXiv:2202.04053*, 2022. 2, 4, 5, 6

[6] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *arXiv preprint arXiv:2204.14217*, 2022. 8

[7] Wan-Cyuan Fan, Yen-Chun Chen, DongDong Chen, Yu Cheng, Lu Yuan, and Yu-Chiang Frank Wang. Frido: Feature pyramid diffusion for complex scene image synthesis. *arXiv preprint arXiv:2208.13753*, 2022. 2, 3

[8] Stanislav Frolov, Prateek Bansal, Jörn Hees, and Andreas Dengel. Dt2i: Dense text-to-image generation from region descriptions. In *Artificial Neural Networks and Machine Learning–ICANN 2022: 31st International Conference on Artificial Neural Networks, Bristol, UK, September 6–9, 2022, Proceedings, Part II*, pages 395–406. Springer, 2022. 3

[9] Stanislav Frolov, Avneesh Sharma, Jörn Hees, Tushar Karayil, Federico Raue, and Andreas Dengel. Attrlostgan: attribute controlled image synthesis from reconfigurable layout and style. In *DAGM German Conference on Pattern Recognition*, pages 361–375. Springer, 2021. 2, 3, 7

[10] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. *arXiv preprint arXiv:2203.13131*, 2022. 3, 8

[11] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 4, 5, 7

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5

[14] Tobias Hinz, Stefan Heinrich, and Stefan Wermter. Generating multiple objects at spatially distinct locations. *arXiv preprint arXiv:1901.00686*, 2019. 3

[15] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 5, 8

[16] Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee. Inferring semantic layout for hierarchical text-to-image synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7986–7994, 2018. 3

[17] Xun Huang, Arun Mallya, Ting-Chun Wang, and Ming-Yu Liu. Multimodal conditional image synthesis with product-of-experts gans. In *European Conference on Computer Vision*, pages 91–109. Springer, 2022. 3

[18] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1219–1228, 2018. 5

[19] Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Text-to-image generation grounded by fine-grained user attention. In *WACV*, pages 237–246, 2021. 3

[20] Bowen Li, Xiaojuan Qi, Philip HS Torr, and Thomas Lukasiewicz. Image-to-image translation with text guidance. *arXiv preprint arXiv:2002.05235*, 2020. 3

[21] Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, and Jianfeng Gao. Object-driven text-to-image synthesis via adversarial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12174–12182, 2019. 3

[22] Yandong Li, Yu Cheng, Zhe Gan, Licheng Yu, Liqiang Wang, and Jingjing Liu. Bachgan: High-resolution image synthesis from salient object layout. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8365–8374, 2020. 2, 3, 7

[23] Zejian Li, Jingyu Wu, Immanuel Koh, Yongchuan Tang, and Lingyun Sun. Image synthesis from layout with locality-aware mask adaption. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13819–13828, 2021. 2, 3, 7

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2, 4, 5, 6, 7, 8

[25] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*, 2022. 5

[26] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[27] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11410–11420, 2022. 5

[28] Dario Pavllo, Aurelien Lucchi, and Thomas Hofmann. Controlling style and semantics in weakly-supervised image generation. In *European conference on computer vision*, pages 482–499. Springer, 2020. 3

[29] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *European conference on computer vision*, pages 647–664. Springer, 2020. 3

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 3

[31] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1, 2, 3

[32] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 4, 5, 8

[33] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR, 2016. 3

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 2, 3, 4, 8

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 3

[36] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 1, 3

[37] Christoph Schuhmann, Romain Beaumont, Cade W Gordon, Ross Wightman, Theo Coombes, Aarush Katta, Clayton Mullis, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 8

[38] Wei Sun and Tianfu Wu. Image synthesis from reconfigurable layout and style. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10531–10540, 2019. 2, 3, 7

[39] Tristan Sylvain, Pengchuan Zhang, Yoshua Bengio, R Devon Hjelm, and Shikhar Sharma. Object-centric image generation from layouts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021. 5

[40] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022. 4

[41] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022. 4, 5, 8

[42] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018. 3

[43] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang. Unitab: Unifying text and box outputs for grounded vision-language modeling. In *European Conference on Computer Vision*, pages 521–539. Springer, 2022. 4

[44] Zuopeng Yang, Daqing Liu, Chaoyue Wang, Jie Yang, and Dacheng Tao. Modeling image composition for complex scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7764–7773, 2022. 2, 3

[45] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*. 1, 2, 3, 4, 5, 8

[46] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 833–842, 2021. 3, 8

[47] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017. 3

[48] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1947–1962, 2018. 3

[49] Bo Zhao, Lili Meng, Weidong Yin, and Leonid Sigal. Image generation from layout. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8584–8593, 2019. 2, 3, 5, 6, 7

[50] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Towards language-free training for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17907–17917, 2022. 8