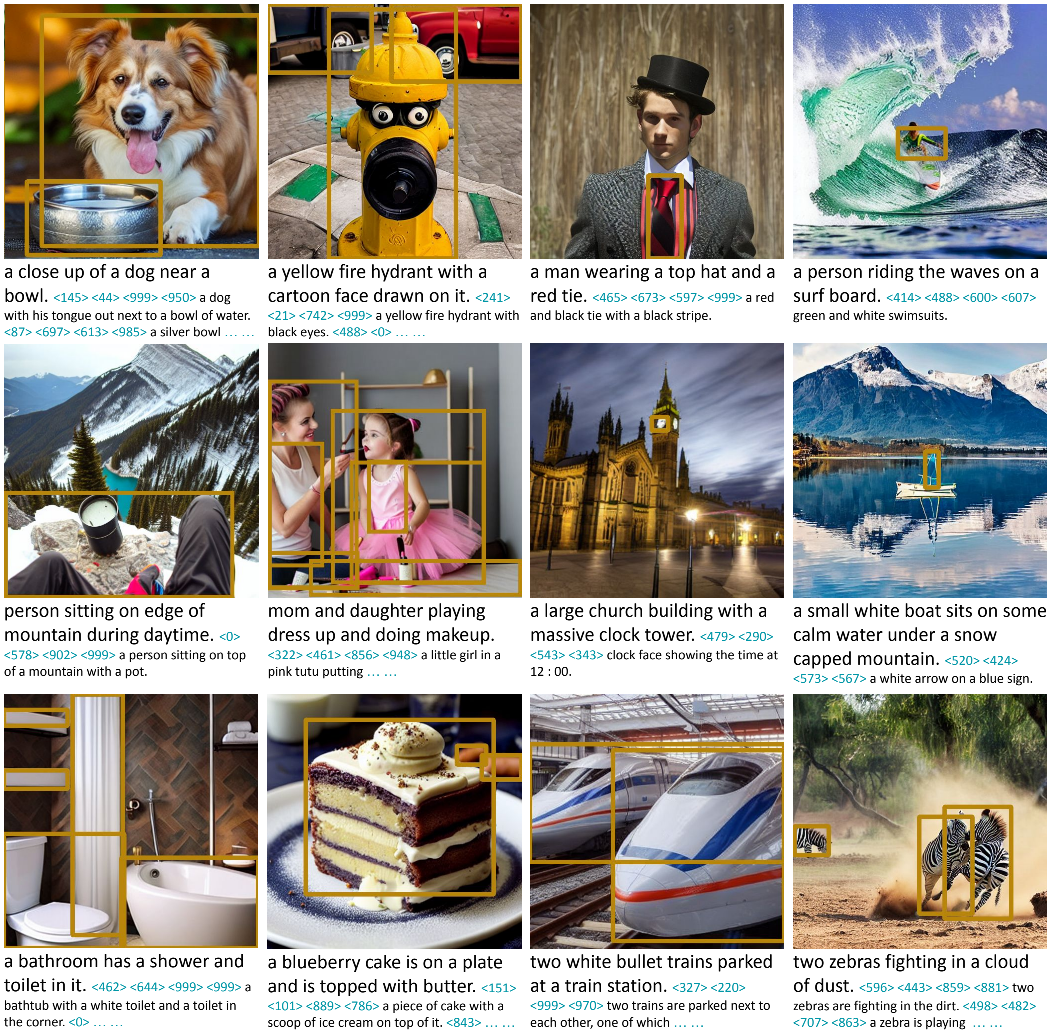
ReCo: Region-Controlled Text-to-Image Generation
(Supplementary Material)



a close up of a dog near a bowl. <145> <44> <999> <950> a dog with his tongue out next to a bowl of water. <87> <697> <613> <985> a silver bowl … …

a yellow fire hydrant with a cartoon face drawn on it. <241> <21> <742> <999> a yellow fire hydrant with black eyes. <488> <0> … …

a man wearing a top hat and a red tie. <465> <673> <597> <999> a red and black tie with a black stripe.

a person riding the waves on a surf board. <414> <488> <600> <607> green and white swimsuits.

person sitting on edge of mountain during daytime. <0> <578> <902> <999> a person sitting on top of a mountain with a pot.

mom and daughter playing dress up and doing makeup. <322> <461> <856> <948> a little girl in a pink tutu putting … …

a large church building with a massive clock tower. <479> <290> <543> <343> clock face showing the time at 12 : 00.

a small white boat sits on some calm water under a snow capped mountain. <520> <424> <573> <567> a white arrow on a blue sign.

a bathroom has a shower and toilet in it. <462> <644> <999> <999> a bathtub with a white toilet and a toilet in the corner. <0> … …

a blueberry cake is on a plate and is topped with butter. <151> <101> <889> <786> a piece of cake with a scoop of ice cream on top of it. <843> … …

two white bullet trains parked at a train station. <327> <220> <999> <970> two trains are parked next to each other, one of which … …

two zebras fighting in a cloud of dust. <596> <443> <859> <881> two zebras are fighting in the dirt. <498> <482> <707> <863> a zebra is playing … …

Figure A. Example images generated by $\text{ReCo}_{\text{LAION}}$. More examples are in Figures G, H.

A. ReCo with LAION data

In the main paper, we focus on the ReCo model trained on COCO ($\text{ReCo}_{\text{COCO}}$) to standardize the evaluation process. In this section, we present $\text{ReCo}_{\text{LAION}}$ 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 $\text{ReCo}_{\text{LAION}}$-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 dataset[1]. We take the Detic object detector [9] to generate

[1] We 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.

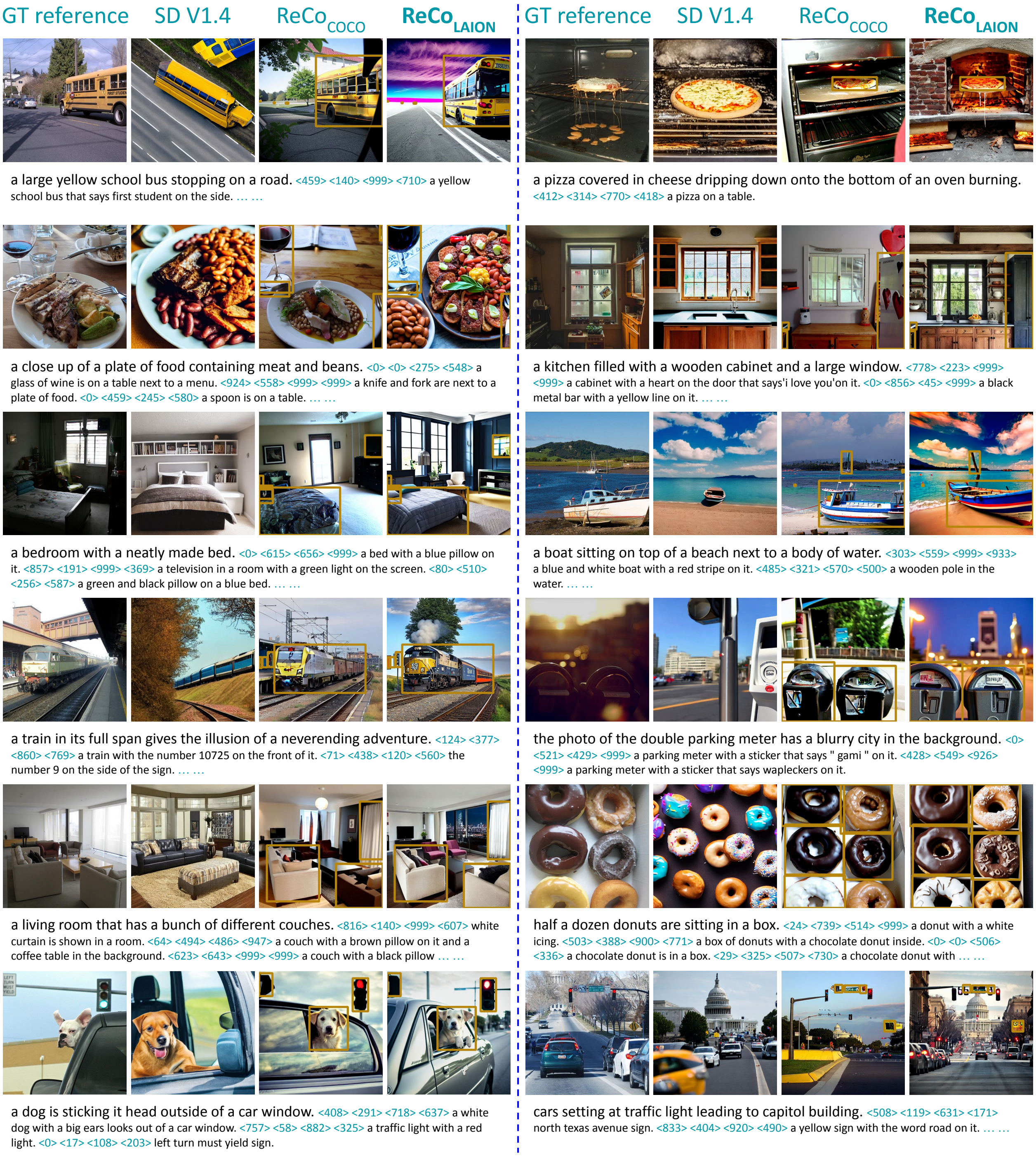
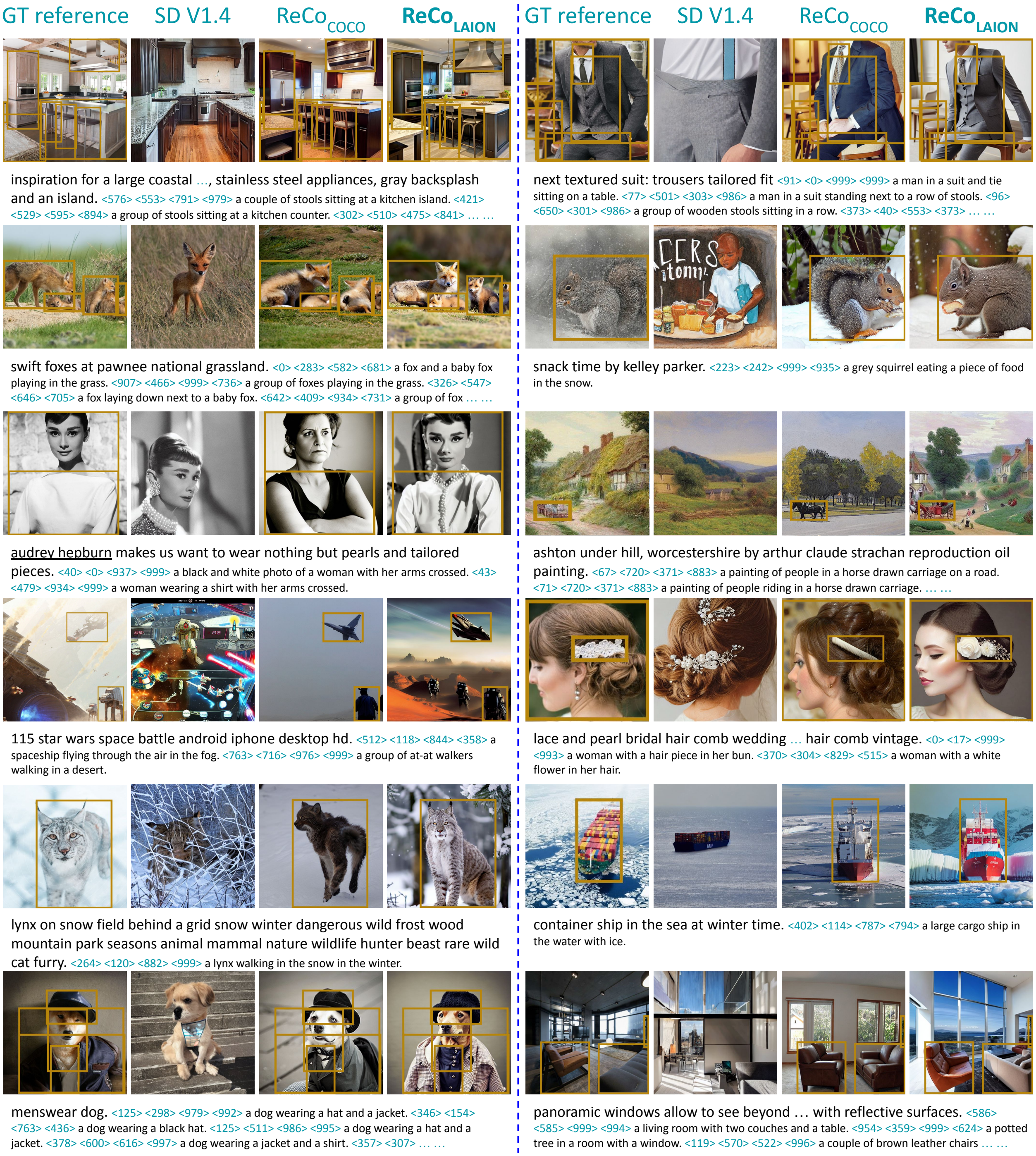

GT reference
SD V1.4
ReCo$_{\text{COCO}}$
ReCo$_{\text{LAION}}$
GT reference
SD V1.4
ReCo$_{\text{COCO}}$
ReCo$_{\text{LAION}}$
a large yellow school bus stopping on a road. <459> <140> <999> <710> a yellow school bus that says first student on the side. … …
a close up of a plate of food containing meat and beans. <0> <0> <275> <548> a glass of wine is on a table next to a menu. <924> <558> <999> <999> a knife and fork are next to a plate of food. <0> <459> <245> <580> a spoon is on a table. … …
a bedroom with a neatly made bed. <0> <615> <656> <999> a bed with a blue pillow on it. <857> <191> <999> <369> a television in a room with a green light on the screen. <80> <510> <256> <587> a green and black pillow on a blue bed. … …
a train in its full span gives the illusion of a neverending adventure. <124> <377> <860> <769> a train with the number 10725 on the front of it. <71> <438> <120> <560> the number 9 on the side of the sign. … …
a living room that has a bunch of different couches. <816> <140> <999> <607> white curtain is shown in a room. <64> <494> <486> <947> a couch with a brown pillow on it and a coffee table in the background. <623> <643> <999> <999> a couch with a black pillow … …
a dog is sticking it head outside of a car window. <408> <291> <718> <637> a white dog with a big ears looks out of a car window. <757> <58> <882> <325> a traffic light with a red light. <0> <17> <108> <203> left turn must yield sign.
a pizza covered in cheese dripping down onto the bottom of an oven burning. <412> <314> <770> <418> a pizza on a table.
a kitchen filled with a wooden cabinet and a large window. <778> <223> <999> <999> a cabinet with a heart on the door that says'i love you'on it. <0> <856> <45> <999> a black metal bar with a yellow line on it. … …
a boat sitting on top of a beach next to a body of water. <303> <559> <999> <933> a blue and white boat with a red stripe on it. <485> <321> <570> <500> a wooden pole in the water. … …
the photo of the double parking meter has a blurry city in the background. <0> <521> <429> <999> a parking meter with a sticker that says " gami " on it. <428> <549> <926> <999> a parking meter with a sticker that says wapleckers on it.
half a dozen donuts are sitting in a box. <24> <739> <514> <999> a donut with a white icing. <503> <388> <900> <771> a box of donuts with a chocolate donut inside. <0> <0> <506> <336> a chocolate donut is in a box. <29> <325> <507> <730> a chocolate donut with … …
cars setting at traffic light leading to capitol building. <508> <119> <631> <171> north texas avenue sign. <833> <404> <920> <490> a yellow sign with the word road on it. … …

Figure B. Qualitative results on LVIS [2]. Zoomed-in version is in Figure G.

the object region predictions. We use a confidence threshold of 0.5 and filter out small boxes with a size smaller than $0.03 \times W \times H$. Following the setting for ReCo$_{\text{COCO}}$, 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 ReCo$_{\text{COCO}}$ and ReCo$_{\text{LAION}}$ show strong region-controlled T2I generation capabilities. Compared with ReCo$_{\text{COCO}}$, ReCo$_{\text{LAION}}$-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.



inspiration for a large coastal …, stainless steel appliances, gray backsplash and an island. <576> <553> <791> <979> a couple of stools sitting at a kitchen island. <421> <529> <595> <894> a group of stools sitting at a kitchen counter. <302> <510> <475> <841> … …

next textured suit: trousers tailored fit <91> <0> <999> <999> a man in a suit and tie sitting on a table. <77> <501> <303> <986> a man in a suit standing next to a row of stools. <96> <650> <301> <986> a group of wooden stools sitting in a row. <373> <40> <553> <373> … …

swift foxes at pawnee national grassland. <0> <283> <582> <681> a fox and a baby fox playing in the grass. <907> <466> <999> <736> a group of foxes playing in the grass. <326> <547> <646> <705> a fox laying down next to a baby fox. <642> <409> <934> <731> a group of fox … …

snack time by kelley parker. <223> <242> <999> <935> a grey squirrel eating a piece of food in the snow.

audrey hepburn makes us want to wear nothing but pearls and tailored pieces. <40> <0> <937> <999> a black and white photo of a woman with her arms crossed. <43> <479> <934> <999> a woman wearing a shirt with her arms crossed.

ashton under hill, worcestershire by arthur claude strachan reproduction oil painting. <67> <720> <371> <883> a painting of people in a horse drawn carriage on a road. <71> <720> <371> <883> a painting of people riding in a horse drawn carriage. … …

115 star wars space battle android iphone desktop hd. <512> <118> <844> <358> a spaceship flying through the air in the fog. <763> <716> <976> <999> a group of at-at walkers walking in a desert.

lace and pearl bridal hair comb wedding … hair comb vintage. <0> <17> <999> <993> a woman with a hair piece in her bun. <370> <304> <829> <515> a woman with a white flower in her hair.

lynx on snow field behind a grid snow winter dangerous wild frost wood mountain park seasons animal mammal nature wildlife hunter beast rare wild cat furry. <264> <120> <882> <999> a lynx walking in the snow in the winter.

container ship in the sea at winter time. <402> <114> <787> <794> a large cargo ship in the water with ice.

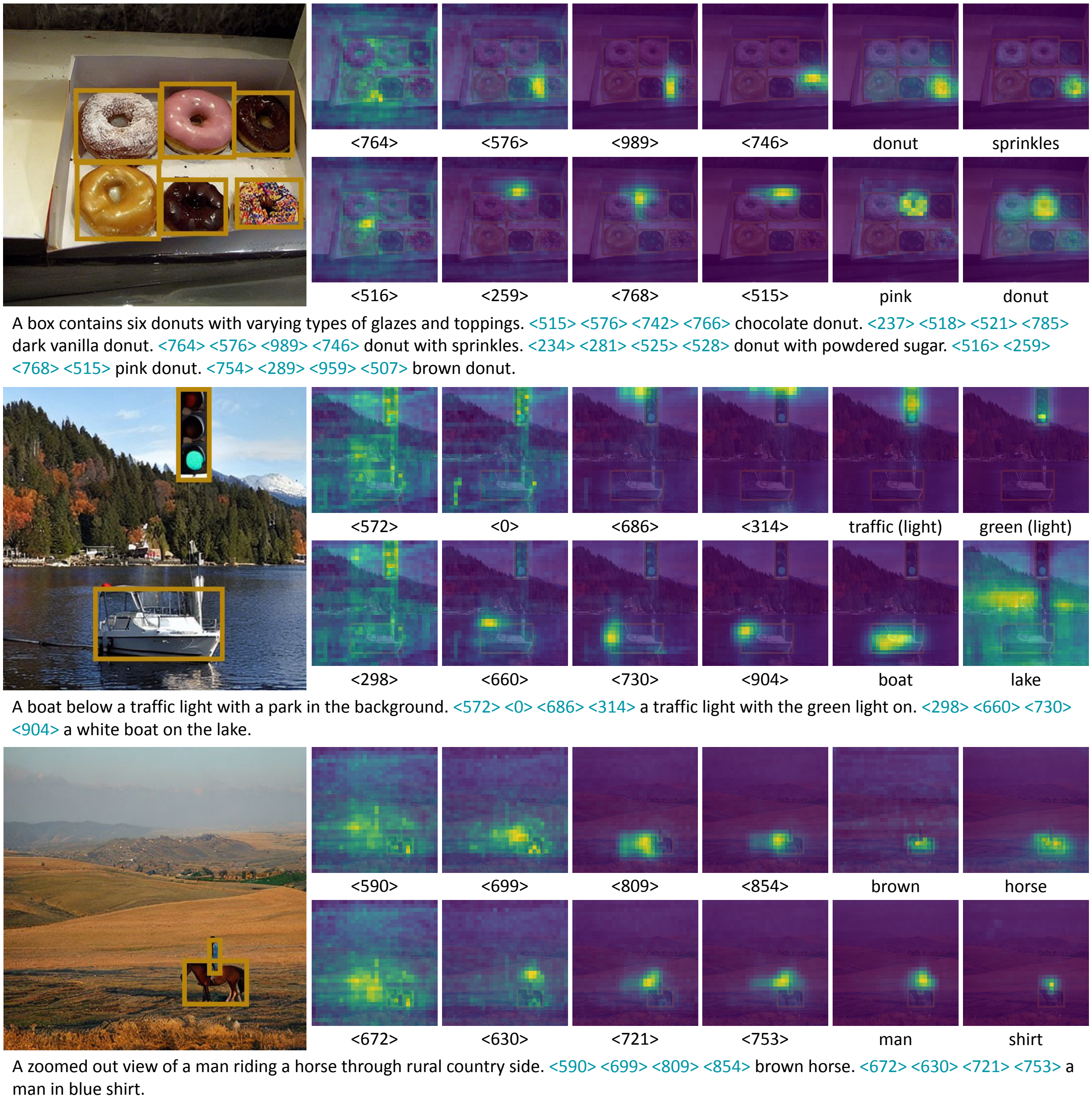
menswear dog. <125> <298> <979> <992> a dog wearing a hat and a jacket. <346> <154> <763> <436> a dog wearing a black hat. <125> <511> <986> <995> a dog wearing a hat and a jacket. <378> <600> <616> <997> a dog wearing a jacket and a shirt. <357> <307> … …

panoramic windows allow to see beyond … with reflective surfaces. <586> <585> <999> <994> a living room with two couches and a table. <954> <359> <999> <624> a potted tree in a room with a window. <119> <570> <522> <996> a couple of brown leather chairs … …

Figure C. Qualitative results on prompts selected from LAION-Aesthetics [5]. Zoomed-in version is in Figure H.

ReCo$_{\text{LAION}}$ 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 ReCo$_{\text{LAION}}$ with ReCo$_{\text{COCO}}$ on LVIS [2]. The "COCO Image" column indicates if the COCO image style is seen during ReCo finetuning. Automatic metrics show that ReCo$_{\text{COCO}}$ 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, ReCo$_{\text{COCO}}$ has seen COCO images during ReCo training, leading to better FID scores. Qualitatively, ReCo$_{\text{LAION}}$-generated images show comparable, if not better visual qualities than ReCo$_{\text{COCO}}$. Overall, both ReCo model variants significantly



A box contains six donuts with varying types of glazes and toppings. <515> <576> <742> <766> chocolate donut. <237> <518> <521> <785> dark vanilla donut. <764> <576> <989> <746> donut with sprinkles. <234> <281> <525> <528> donut with powdered sugar. <516> <259> <768> <515> pink donut. <754> <289> <959> <507> brown donut.

A boat below a traffic light with a park in the background. <572> <0> <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 $\text{ReCo}_{\text{COCO}}$ cross-attention maps between visual latent and text embedding (on both text and position tokens).

Method	COCO Image	Object Acc. (↑)	SceneFID (↓)	FID (↓)
Real Images	-	42.00	-	-
SD V1.4	✗	7.88	40.62	23.74
$\text{ReCo}_{\text{COCO}}$	✓	**23.42**	**10.08**	**17.73**
$\text{ReCo}_{\text{LAION}}$	✗	19.38	19.48	21.99

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 $\tau_\theta(y(P,T))$. 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

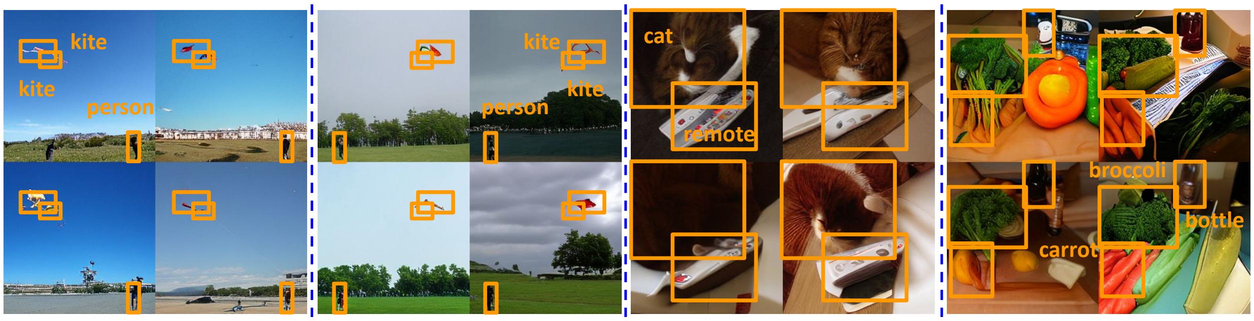


Figure E. ReCo$_{\text{OFA}}$ based on the auto-regressive T2I model OFA$_{\text{Large}}$ [7].

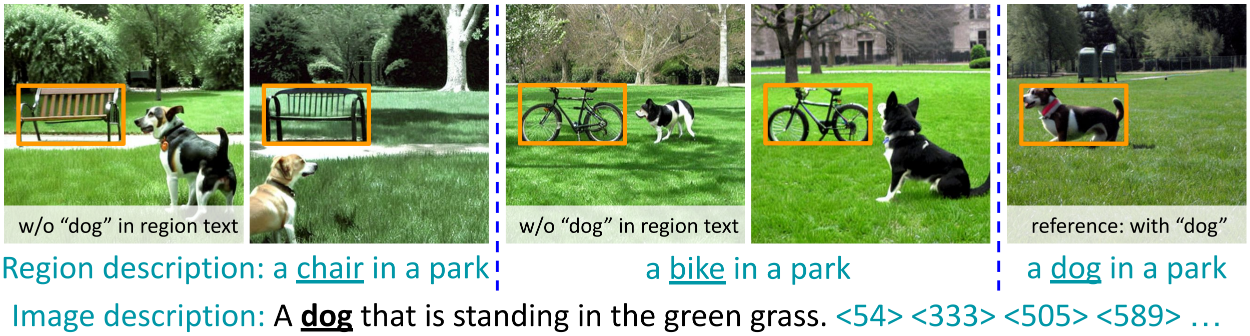

w/o "dog" in region text
w/o "dog" in region text
reference: with "dog"
Region description: a chair in a park
a bike in a park
a dog in a park
Image description: A dog that is standing in the green grass. <54> <333> <505> <589> …

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

regional descriptions, *e.g.*, the "green light" in the "traffic light." We hypothesize that the observed "coarse-to-fine-grained localization" pattern might be related to the causal masking in the text encoder, as early position tokens do not have the complete region information.

C. Supplementary Discussions

Generality beyond diffusion. In the main paper, we present ReCo based on the diffusion-based Stable Diffusion model [4]. To better understand if the core idea could generalize to other T2I systems, we conduct experiments with ReCo$_{\text{OFA}}$, a variation that uses an open-sourced auto-regressive T2I model OFA$_{\text{Large}}$ [7] as the generation backbone. We empirically observe that ReCo's capabilities and findings generalize well to auto-regressive T2I models (*cf*., AP/AP$_{50}$/FID: ReCo$_{\text{OFA}}$ 13.1/24.2/9.10 *vs*. OFA$_{\text{Large}}$ 0.8/2.5/11.82). Specifically, ReCo's region-controlled generation not only provides the desired region controllability but also improves the image generation quality. Figure E shows the results of ReCo$_{\text{OFA}}$'s region-controlled T2I generation.

"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.

Method	AP	AP$_{50}$	Object Acc.	SceneFID	FID
ReCo	32.0	52.4	62.42	6.51	7.36
ReCo$_{\text{Position Word}}$	2.3	7.5	42.02	15.54	8.82
ReCo$_{\text{Relation Word}}$	1.5	4.8	43.99	13.98	9.50

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

ReCo$_{\text{Relation Word}}$ with relationship words. In addition to the position text words used in ReCo$_{\text{Position Word}}$, ReCo$_{\text{Relation Word}}$ 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$_{\text{Position Word}}$, while the performance remains lower than ReCo with position tokens. Therefore, we use ReCo$_{\text{Position Word}}$ as the reference model in the main paper and propose that position tokens could be inherently more concise and accurate for spatial controllability.

References

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

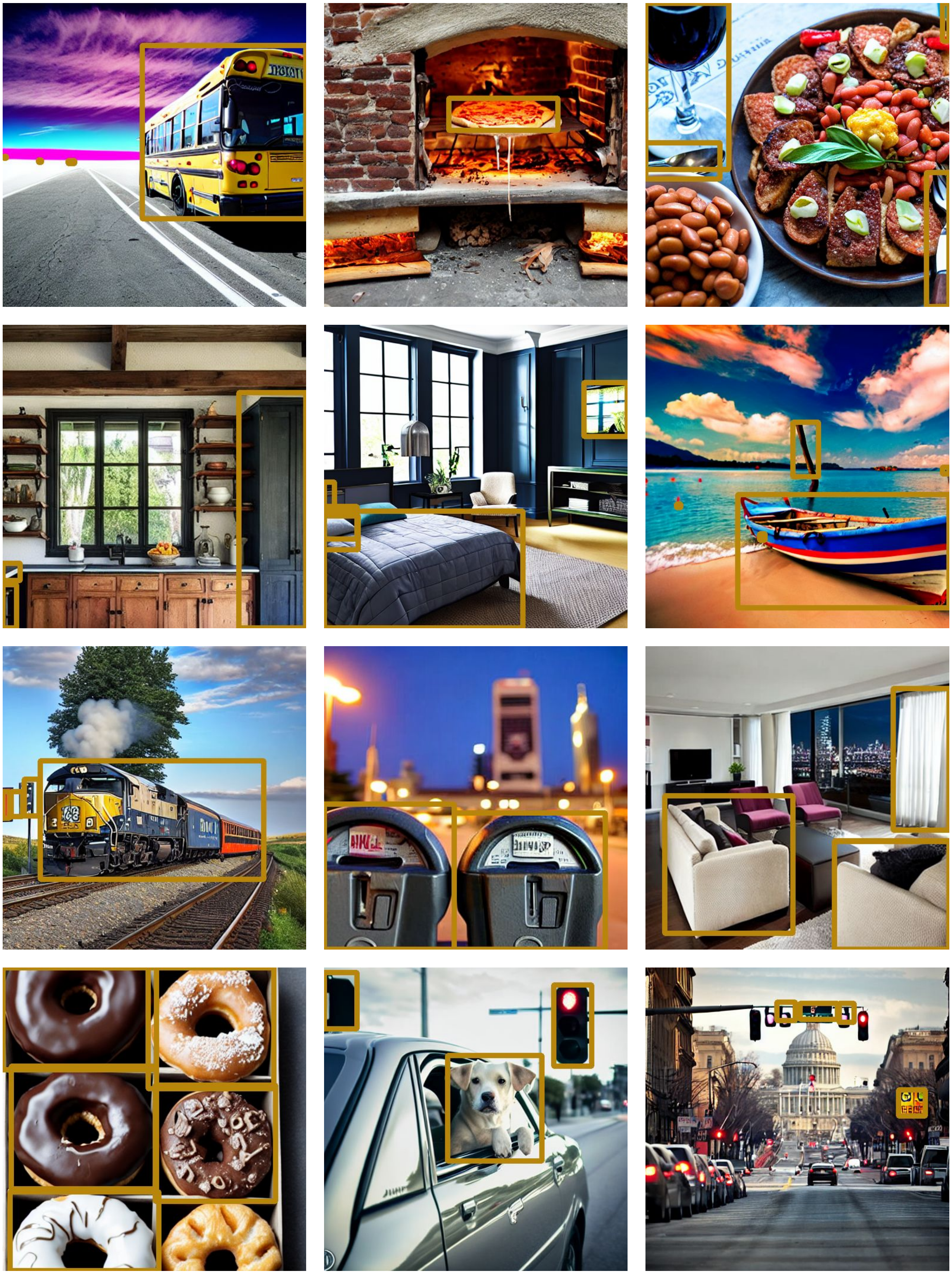


Figure G. Zooming in ReCo$_{\text{LAION}}$-generated images shown in Figure B.

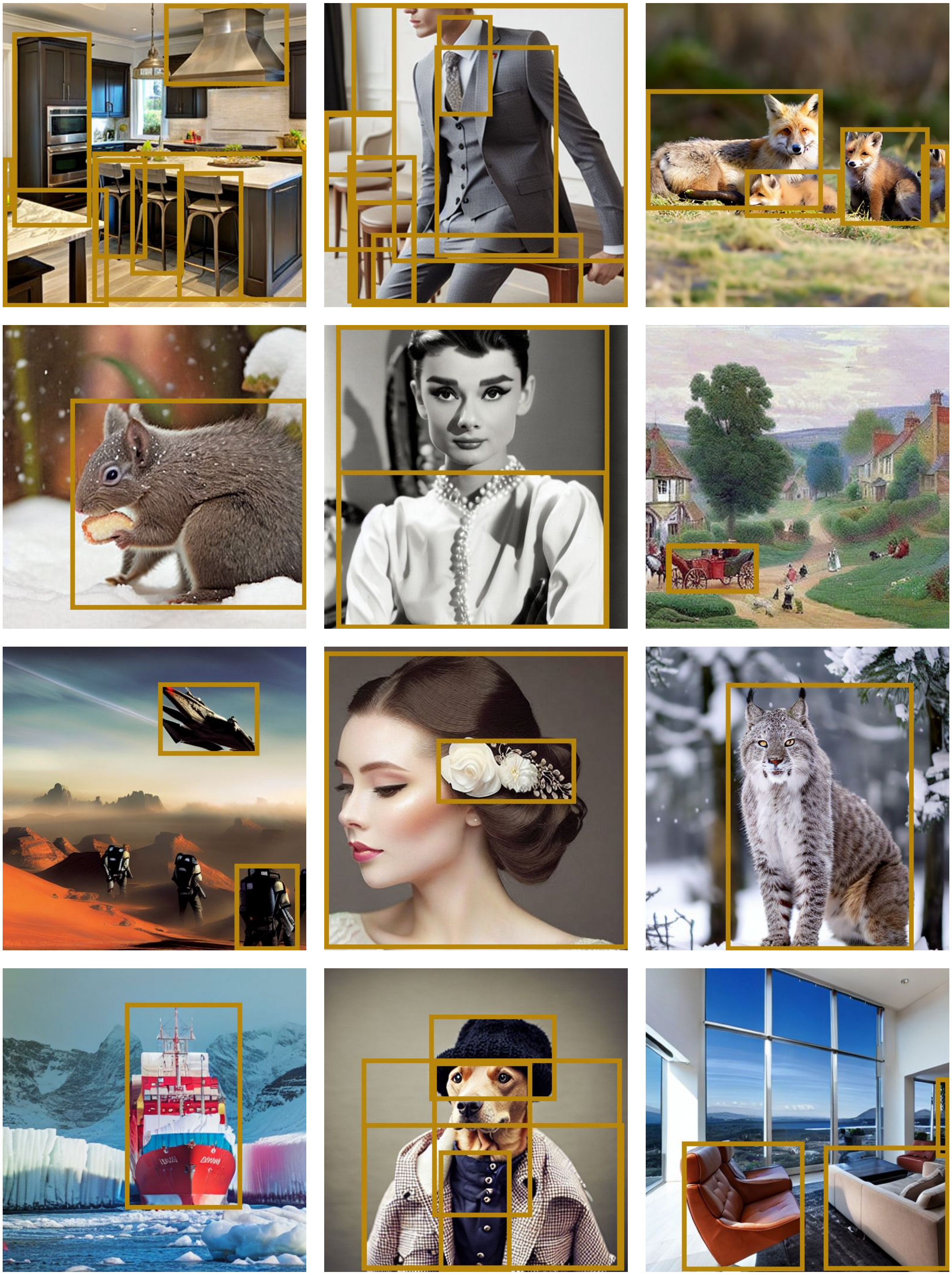


Figure H. Zooming in ReCo$_{\text{LAION}}$-generated images shown in Figure C.

[2] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 2, 3, 4

[3] Yash Kant, Dhruv Batra, Peter Anderson, Alex Schwing, Devi Parikh, Jiasen Lu, and Harsh Agrawal. Spatially aware multimodal transformers for textvqa. In *ECCV*, 2020. 5

[4] 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, 5

[5] 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. 1, 2, 3

[6] 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. 2

[7] 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. 5

[8] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *ECCV*, 2018. 5

[9] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, pages 350–368. Springer, 2022. 1