APPENDIX

A PRELIMINARY

Our model is based on Stable Diffusion v1.4 Rombach et al. (2022), a Latent Diffusion model (LDM) that applies the diffusion process in a latent space. Specifically, an input image x is encoded into the latent space using a pretrained autoencoder $z = \mathcal{E}(x), \hat{x} = \mathcal{D}(z)$ (with an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$). Then the denoising process is achieved by training a denoiser $\epsilon_\theta\left(z_t, t, f_c\right)$ that predicts the added noise following:

$$\min_\theta E_{z_0, \epsilon \sim \mathcal{N}(0,1), t \sim \mathrm{U}(1,T)} \left\|\epsilon - \varepsilon_\theta\left(z_t, t, f_c\right)\right\|_2^2, \tag{8}$$

where f_c is the embedding of the condition (such as a prompt) and z_t is the latent noise at timestamp t.

B TRAINING/INFERENCE DETAILS

Our model is trained on 4 NVIDIA A100 GPUs for 100k steps with a batch size of 14 and a learning rate of 5×10^{-5}. During training, we randomly drop reference image embedding and text embedding both at the rate of 10%. We decently rank the area of the boxes per images, and set the max number of grounding boxes to be 10 with the largest areas. During inference, we set classifier-free guidance(CFG) (Ho & Salimans, 2022) as 3.

C DETAILS ABOUT DATA COLLECTION

For each reference image, we use the segmentation mask to mask out the background and get the background-free reference object. In inference stage, we use SAM (Kirillov et al., 2023) to get the mask of the reference object, and get the background-free reference object.

D DETAILS ABOUT USER STUDY

Our user study is based on DreamBench, with full 30 objects and 25 prompts. We randomly generated layouts, and use them in the generation. In the user study, given the layout, the reference object, the text prompt, the result of our method and a random-selected baseline method, we request the user to answer the following four questions:

(1) Which generated image do you think that its object is more similar to the input object? Choose between Option A and B.

(2) Which generated image do you think that its object is most likely to be at the right position as the input layout? Choose between Option A and B.

(3) Which generated image do you think is most likely to match the text description? Choose between Option A and B.

(4) Which image do you think has better image quality? Choose between Option A and B.

We received more than 1200 votes from over 530 users. In the experiment, we randomly shuffle the order of baselines to improve the confidence of the user study.

E ADDITIONAL QUALITATIVE RESULTS ON POSE CHANGE

In Fig. 8 we show results about changing the shape of the bounding box. For grounded text-to-image customization, different from traditional text-to-image customization, the pose of the object is jointly influenced by the shape of the bounding box and the model's ability to adapt the object to be harmonious with the background. The model tend to first adapt the object to the bounding box, then make pose adjustments to make object to be harmonious with the background. For instance, in the 1st and 4th row of Fig. 8, given a bounding box with a large or small width/height ratio, the grounded customized generation will generate objects with large pose change to adapt to the bounding box,

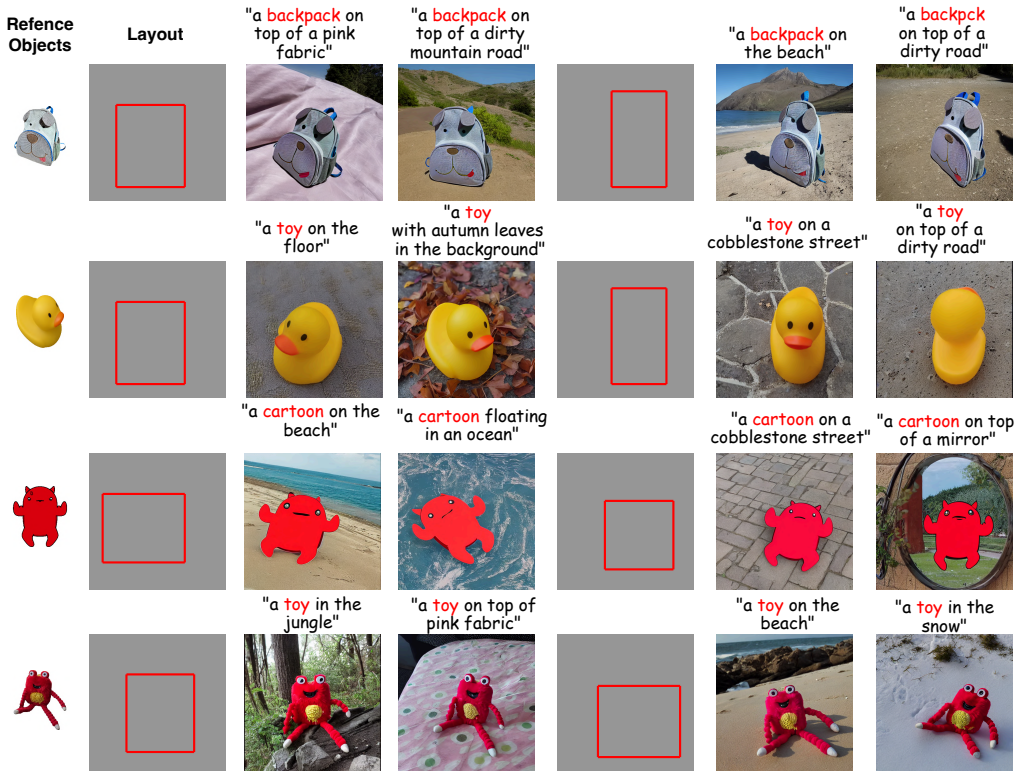

Figure 8: More visual results of our model about layout and pose change: in our model, the pose of the object is influenced by both the shape of the bounding box and the model's ability to adapt to the background. The model tends to first adapt the object into the layout, then adapt the pose to maintain harmonization with the background.

then make pose refinement inside the bounding box. Users can easily conduct the initial manipulation of the object by specifying the desired layout, then the model will automatically adjust the pose of the object to be harmonious with the background. Our model shows both the ability to generate objects with accurate location and the ability to make pose changes to the objects.

F Analysis on Grounding Circumstance

We also show qualitative results under the consumption that no layout is provided by the users. From the results, we can see that: Our model also supports text-to-image generation, layout-to-image generation, and personalized text-to-image generation tasks.

- As shown in Fig. 9, if the bounding box is set to be $[x1, y1, x2, y2] = [0, 0, 0, 0]$, the model will degrade into simpler text-to-image generation task, since the corresponding grounding tokens are set to be all-zero, and the model also loses the grounding ability.

- As shown in Fig. 10, if no reference object as input, and all the layouts rely on the input text entity to generate, then the model will degrade into layout-guided text-to-image generation task.

- If randomly assigned the bounding box of the reference object, our model is equal to the text-to-image personalization task, like previous non-grounding text-to-image customization works.

Figure 9: Our model can also deal with pure text-to-image generation task. When we set the layout $[x1, y1, x2, y2] = [0.0, 0.0, 0.0, 0.0]$, the model will degrade into a simpler text-to-image generation task, since the corresponding grounding tokens are set to be all-zero, and the model also loses the grounding ability.

Figure 10: Our model can also deal with layout-guided text-to-image generation task: when there is no reference image input, the model will degrade into a layout-guided text-to-image generation task.

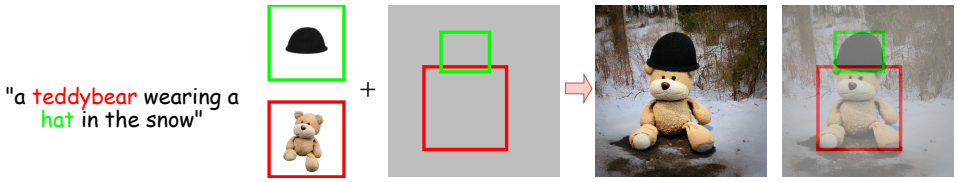

Figure 11: More results about live animals wearing clothes.

G More Results about Object Interaction

As shown in Fig. 11, taking a toy object and a hat as input, our model is able to put the hat on the teddy bear, which shows the model's ability to composite reference objects.

H More Results about pose change under the guidance of prompt

We further show comparison results about pose change under the guidance of prompts in Fig. 12. We select prompts that is relevant to actions and pose change. Previous text-to-image customization models cannot maintain the identity of the reference object(row 2, row 4 and row 5), fail to achieve the prompt action-guided pose change(row 3 and row 4) and maintain text-alignment in certain

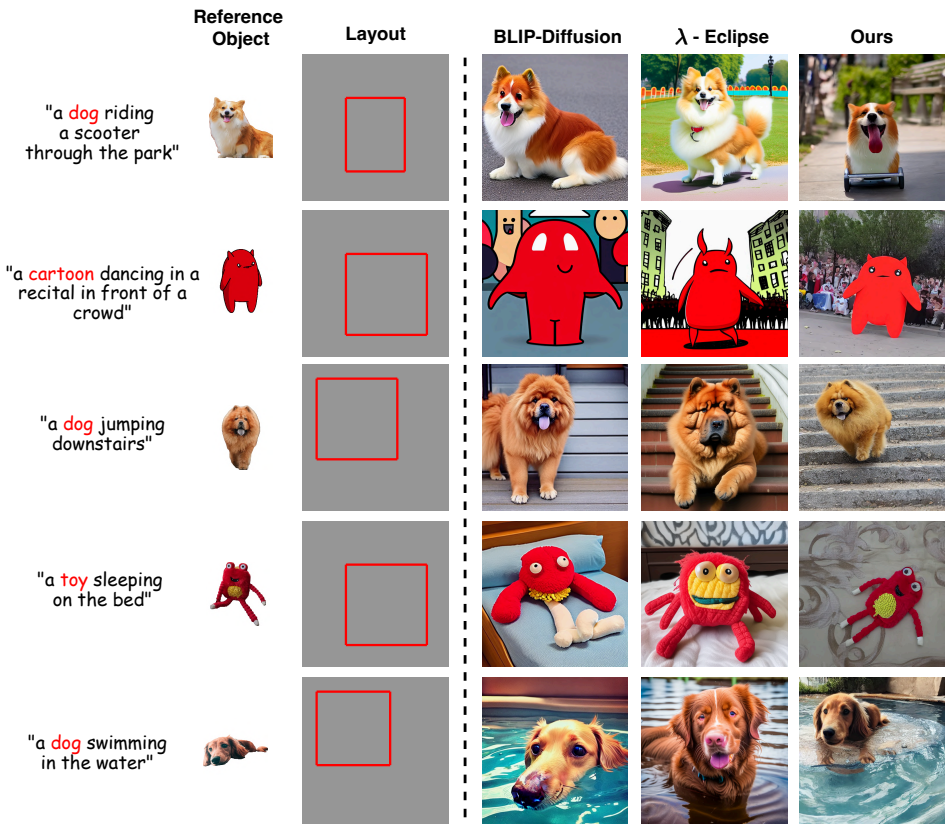

Figure 12: More results about pose change under the guidance of prompt.

Table 6: Comparison with existing methods on Dreambench under layout scale normalization.

	CLIP-T ↑	CLIP-I ↑	DINO-I ↑
SD V1.4 [(Rombach et al., 2022)]	0.3122	0.8413	0.6587
BLIP-Diffusion [(Li et al., 2024)]	0.2824	0.8894	0.7625
ELITE [(Wei et al., 2023)]	0.2461	0.8936	0.7557
Kosmos-G [(Pan et al., 2023)]	0.2864	0.8452	0.6933
lambda-eclipse [(Patel et al., 2024)]	0.2767	0.8973	0.7934
AnyDoor [(Chen et al., 2023b)]	0.2430	0.9062	0.7928
GLIGEN [(Li et al., 2023)]	0.2898	0.8520	0.6890
CustomNet [(Yuan et al., 2023)]	0.2821	0.9103	0.7587
Ours	**0.2911**	**0.9169**	**0.7950**

cases(row 1 and row 3). Our method not only achieve grounded text-to-image customization, but also able to maintain a good balance between identity preservation and text alignment, and can also generate objects with variations in pose.

I Comparison under Layout Scale Normalization

We further conducted experiments to normalize our bounding box scales based on the average size of objects generated by other personalized text-to-image generation methods. We update the comparison results in the Table 6. For non-grounding-based text-to-image customization methods, we used Grounding DINO (Liu et al., 2023b) to detect the bounding box of the target subject by identifying the object name. We then computed the average bounding box area and applied a $\pm 20\%$ variation as the normalized bounding box size. This normalized bounding box size scale was subsequently employed for the grounded text-to-image customization methods(CustomNet (Yuan et al., 2023) and

Ours). The results demonstrate that our method achieves improved CLIP-T, CLIP-I and DINO-I scores, outperforming all baseline personalized text-to-image generation methods and layout-guided text-to-image generation methods in this case.

J Additional Qualitative Results

Here we show more qualitative results. In Fig. 13 we show results on DreamBench and in Fig. 14 and Fig. 15 we show more results about complex background background evaluation on coco validation set.

K Social Impact

GroundingBooth provides a flexible method for users to precisely customize the layout of both foreground and background objects based on user-provided reference subjects and text descriptions without any test-time finetuning. The support for the generation of multi-subjects provides a useful tool for users to generate images using their desired layout. Users can optionally choose reference objects or simple text inputs to generate their desired image, which significantly expands the flexibility in controllable and customized text-to-image generation.

Figure 13: More visual results of our model.

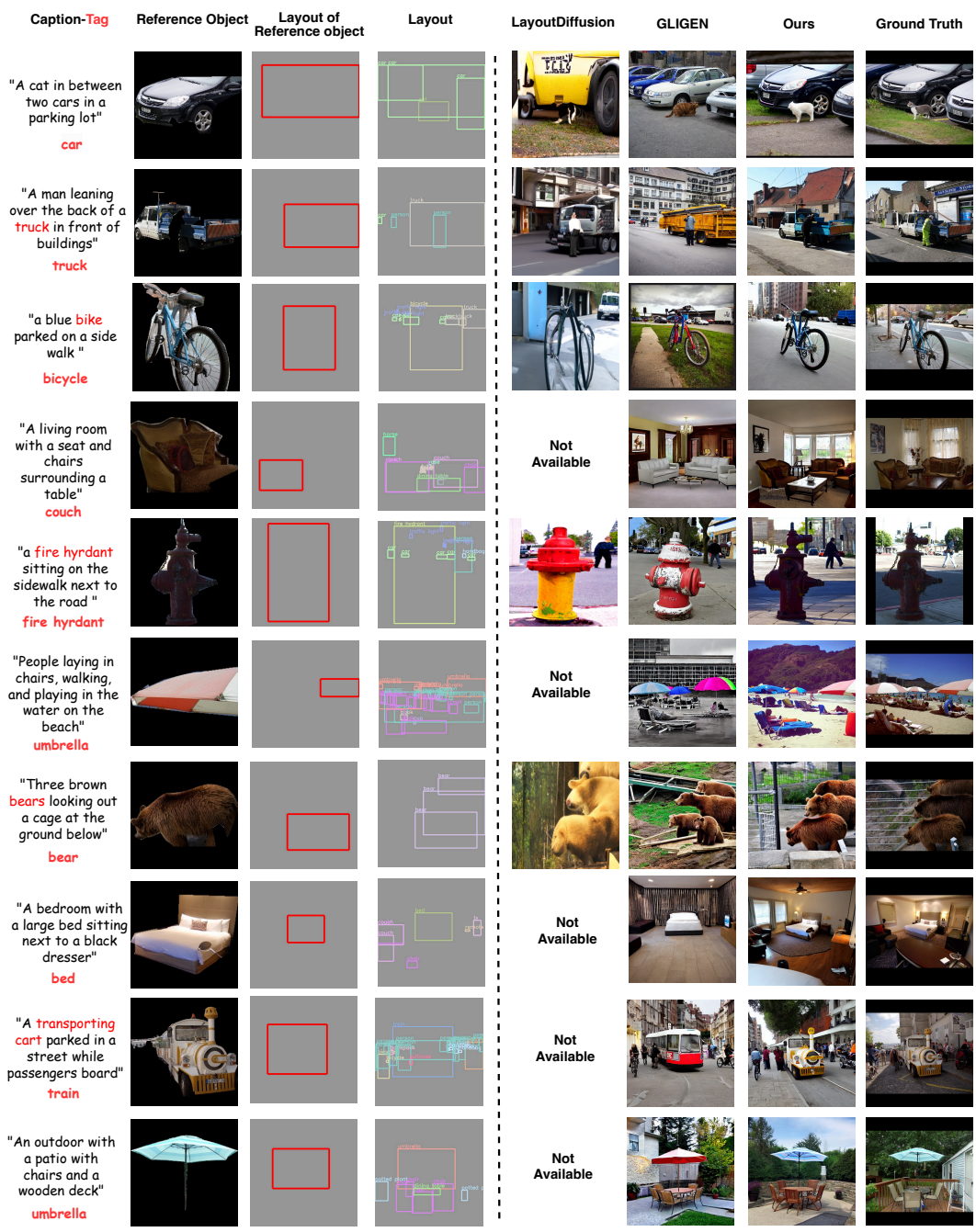

Figure 14: More results on complex scene generation on COCO validation set.

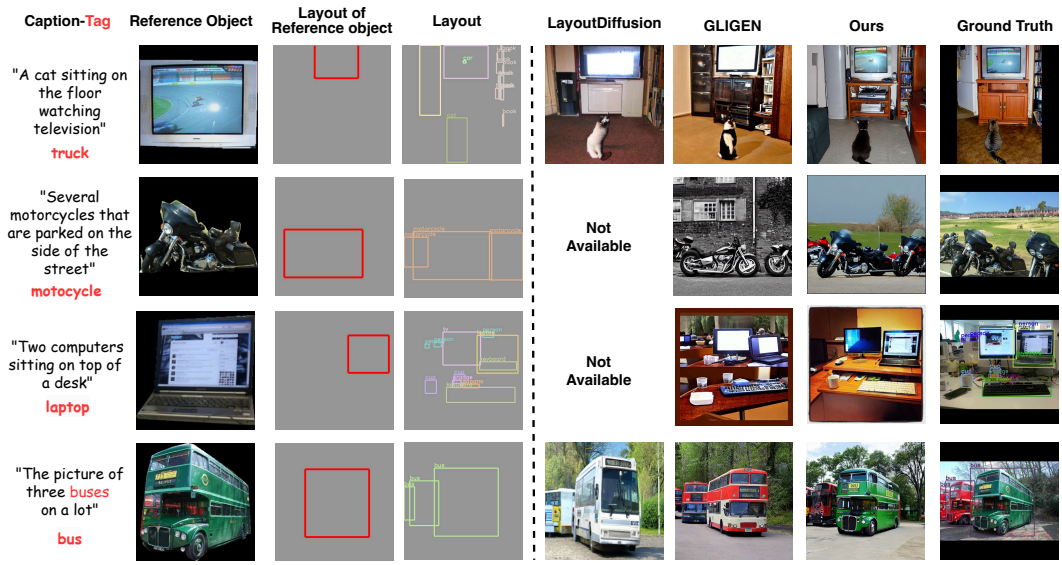

Figure 15: More results on complex scene generation on COCO validation set.