ReGround: Improving Textual and Spatial Grounding at No Cost – Supplementary Material

Phillip Y. Lee Minhyuk Sung

KAIST
{phillip0701,mhsung}@kaist.ac.kr

In this supplementary material, we provide additional details on the evaluation setup (Sec. S.1) and more quantitative comparisons of our ReGround and GLIGEN [3] (Sec. S.2). Moreover, we showcase the effect of ReGround as a backbone of zero-shot layout-guided image generation methods (Sec. S.3) and finally provide extensive qualitative comparisons of Stable Diffusion (SD) [8], GLIGEN [3], and our ReGround (Sec. S.4).

S.1 Details on Evaluation Setup

This section provides further descriptions on the evaluation datasets (Sec. 6.1) and the user study setup (Sec. 6.2) in the main paper

MS-COCO. The validation set of the MS-COCO-2017 dataset [5] consists of 5,000 image-annotation pairs. Since GLIGEN [3] is trained to handle a maximum of 30 bounding boxes per image, we excluded pairs with more than 30 bounding boxes or no bounding boxes, resulting in a total of 4,952 images. For the validation set of the MS-COCO-2014 dataset [5], we randomly sampled 5,000 pairs for evaluation.

NSR-1K-GPT. *Numerical and Spatial Reasoning (NSR-1K)* [1] is a collection of layout-caption pairs designed to assess the numerical and spatial reasoning capabilities of image generation methods. The object labels and bounding boxes are from MS-COCO [5], while the captions are newly annotated based on the spatial relationships and numerical properties of objects. NSR-1K consists of two subsets: *Counting* and *Spatial*. We randomly sampled 1,000 pairs from the Counting set and used all 1,021 pairs from the Spatial set.

User Study. We conducted the user study through Amazon Mechanical Turk using the template displayed in Fig. S1. Based on the text prompt and bounding boxes generated from GPT-4 [6], images were generated by both GLIGEN [3] and our ReGround. Since ReGround aims to resolve the failure cases of GLIGEN, we re-generated both images when the differences between them were minimal (*i.e.*, if the LPIPS value [11] was less than 0.3), resulting in an average of 2.4 iterations per image. Each participant answered 20 questions and 5 vigilance tests.

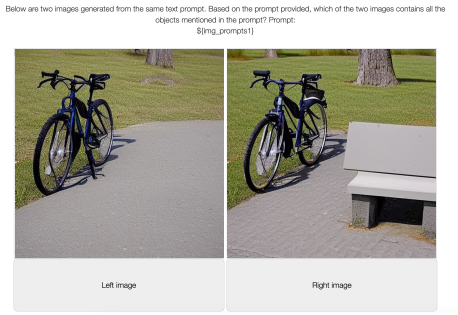
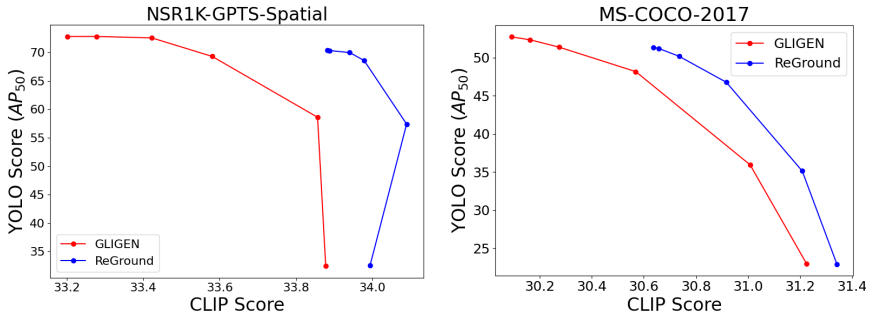


Fig. S1: User study template. In the above example, the text prompt *"A photo of a bicycle and a bench"* was displayed to the respondents.

S.2 Additional Quantitative Comparisons

In addition to Sec. 6.3 of the main paper, this section provides quantitative comparisons between GLIGEN [3] and our ReGround on the Spatial subset of NSR-1K-GPT, and with a different version of Stable Diffusion [8] as the base image diffusion model.

Comparison on NSR-1K-GPT-Spatial. Fig. S2-(a) shows the CLIP score [2] and YOLO score [9] measured on the *Spatial* subset of NSR-1K-GPT. The minimum CLIP score of our ReGround (33.89 at $\gamma = 1.0$) is already higher than GLIGEN's maximum CLIP score (33.88 at $\gamma = 0.1$), indicating that ReGround obtains a significant enhancement in textual grounding while preserving the spatial grounding.



(a) NSR-1K-GPT-Spatial **(SDv1.4)** (b) MS-COCO-2017 **(SDv2.1)**

Fig. S2: Quantitative comparisons (a) on the *Spatial* subset of NSR-1K-GPT and (b) using SDv2.1 as the base image diffusion model. Consistent with the findings from Fig. 6 of the main paper, our ReGround demonstrates improved performance in textual and spatial groundings, as seen by the higher CLIP score [2] for the same range of YOLO score [9].

Results with SDv2.1 as Base Diffusion Model. In Sec. 6 of the main paper, we conducted experiments using the GLIGEN [3] checkpoint based on Stable Diffusion v1.4 **(SDv1.4)**. Additionally, we provide quantitative comparisons with an unofficial GLIGEN checkpoint [4] that was trained with **SDv2.1** as the base image diffusion model. The results, presented in Fig. S2-(b), clearly demonstrate the significant outperformance of our ReGround over GLIGEN.

S.3 More Results with ReGround as Backbone

In addition to Sec. 6.4 of the main paper, we provide qualitative comparisons of different layout-guided generation methods using GLIGEN [3] and our ReGround as backbones, respectively (Fig. S3, S4). The results on BoxDiff [10] and Attention Refocusing [7] illustrate that our network rewiring substantially improves the performance of layout-guided generation methods built upon the GLIGEN framework.

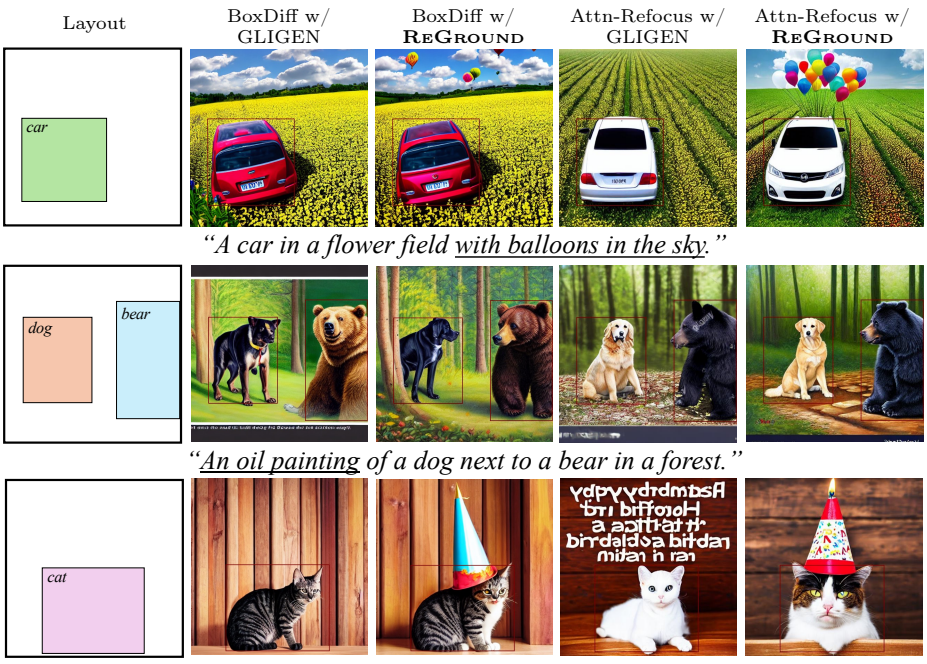


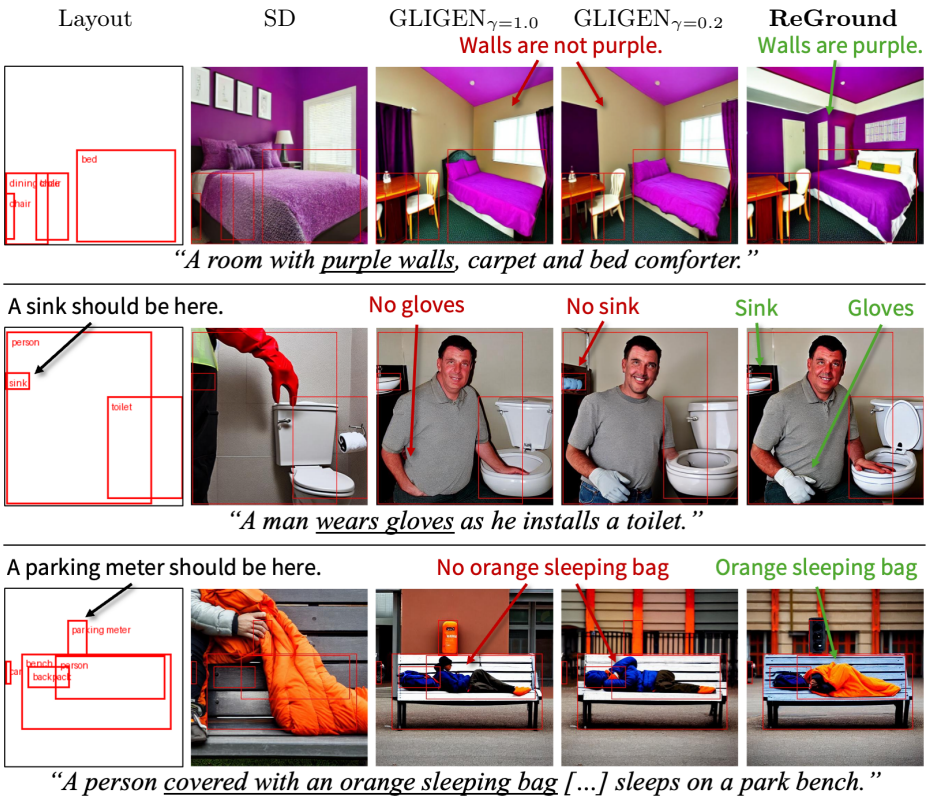
Fig. S3: Comparisons of GLIGEN [3] and our ReGround as a backbone for BoxDiff [10] and Attention Refocusing (Attn-Refocus) [7].

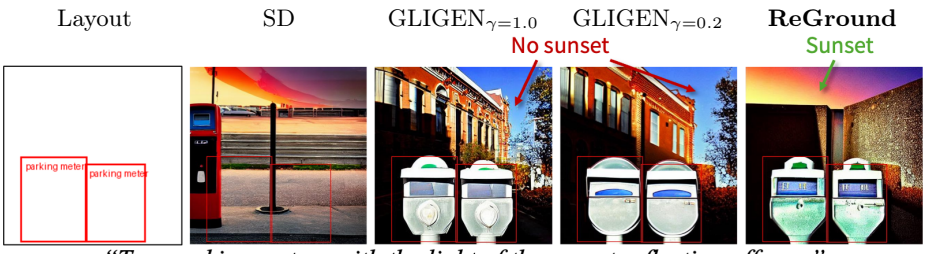
Fig. S4: More comparisons on BoxDiff [10] and Attention Refocusing (Attn-Refocus) [7].

S.4 Additional Qualitative Comparisons

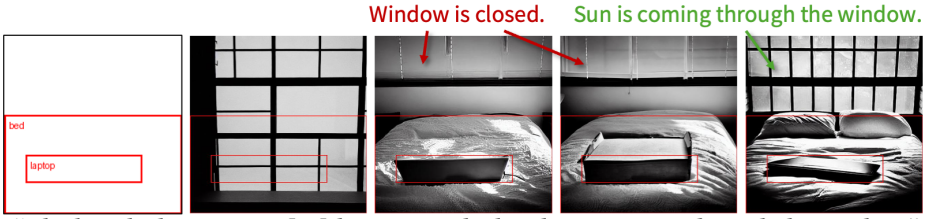
In this section, we provide extensive qualitative comparisons of Stable Diffusion (SD) [8], GLIGEN [3], and our ReGround on layout-guided image generation. Note that $\gamma \in [0, 1]$ denotes the fraction of the initial denoising steps during which gated self-attention is activated, as discussed in Sec. 5.1 of the main paper.

In each row, the input layout is presented in the first column, with the input text prompt displayed below the images. The phrase underlined in each prompt highlights the entity subject to *description omission*, as mentioned in Sec. 4.2 of the main paper. Furthermore, black arrows are used to denote bounding boxes that some methods fail to represent accurately, whereas other methods succeed in doing so precisely. Red arrows signify a failure in either spatial or textual grounding, while green arrows indicate successful grounding of a specific entity.

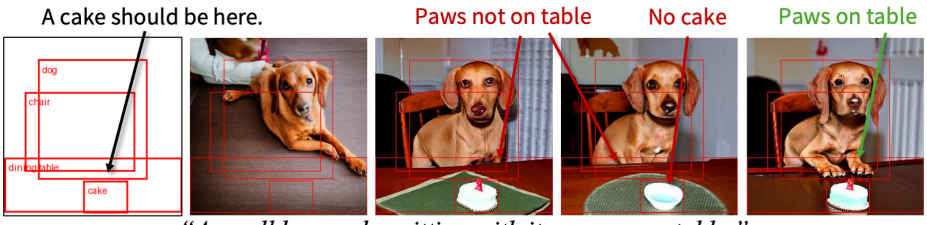




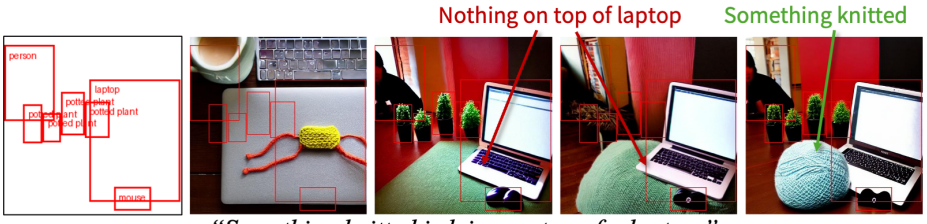
"Two parking meters with the light of the sunset reflecting off one."



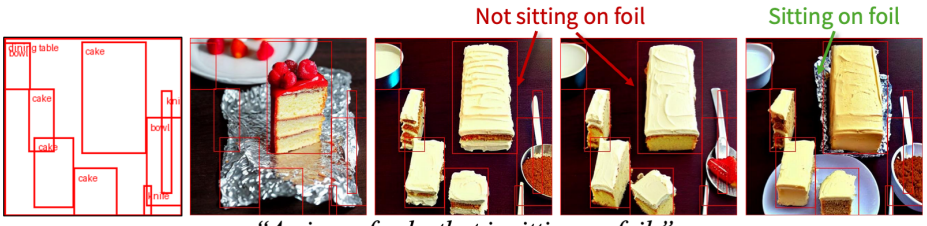
"Black and white picture [...] laying on a bed with sun coming through the window."



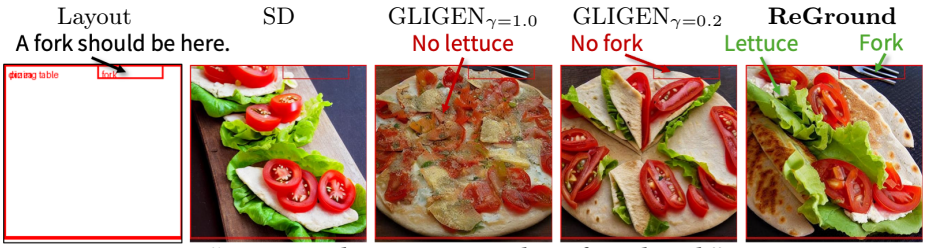
"A small brown dog sitting with its paws on a table."



"Something knitted is lying on top of a laptop."



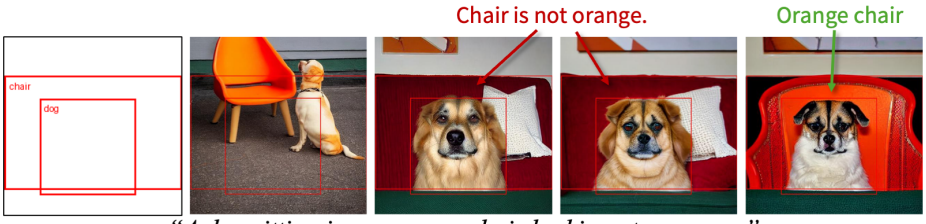
"A piece of cake that is sitting on foil."



"Lettuce and tomatoes on a slice of pita bread."

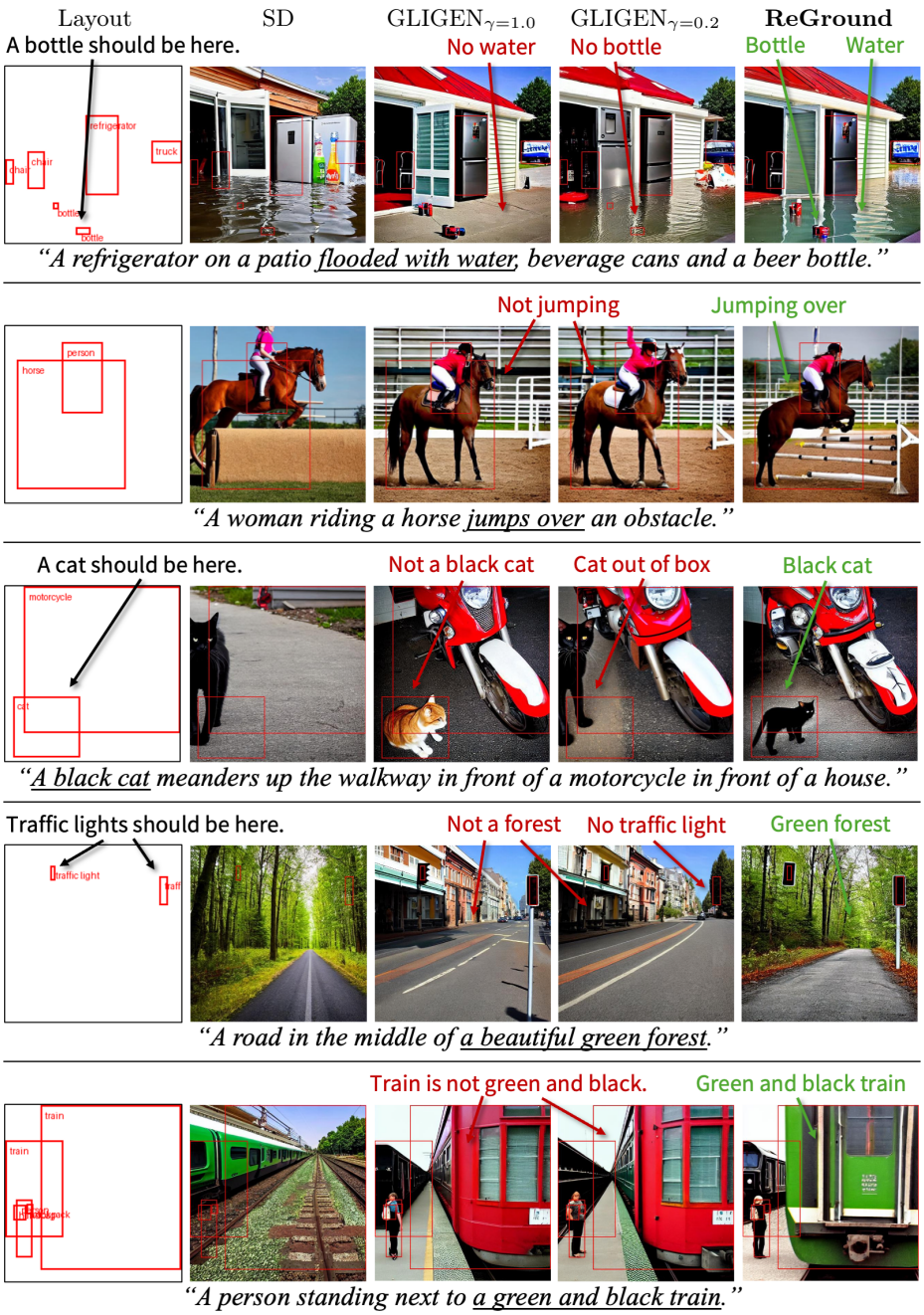
"A restroom with a toilet, sink and window."

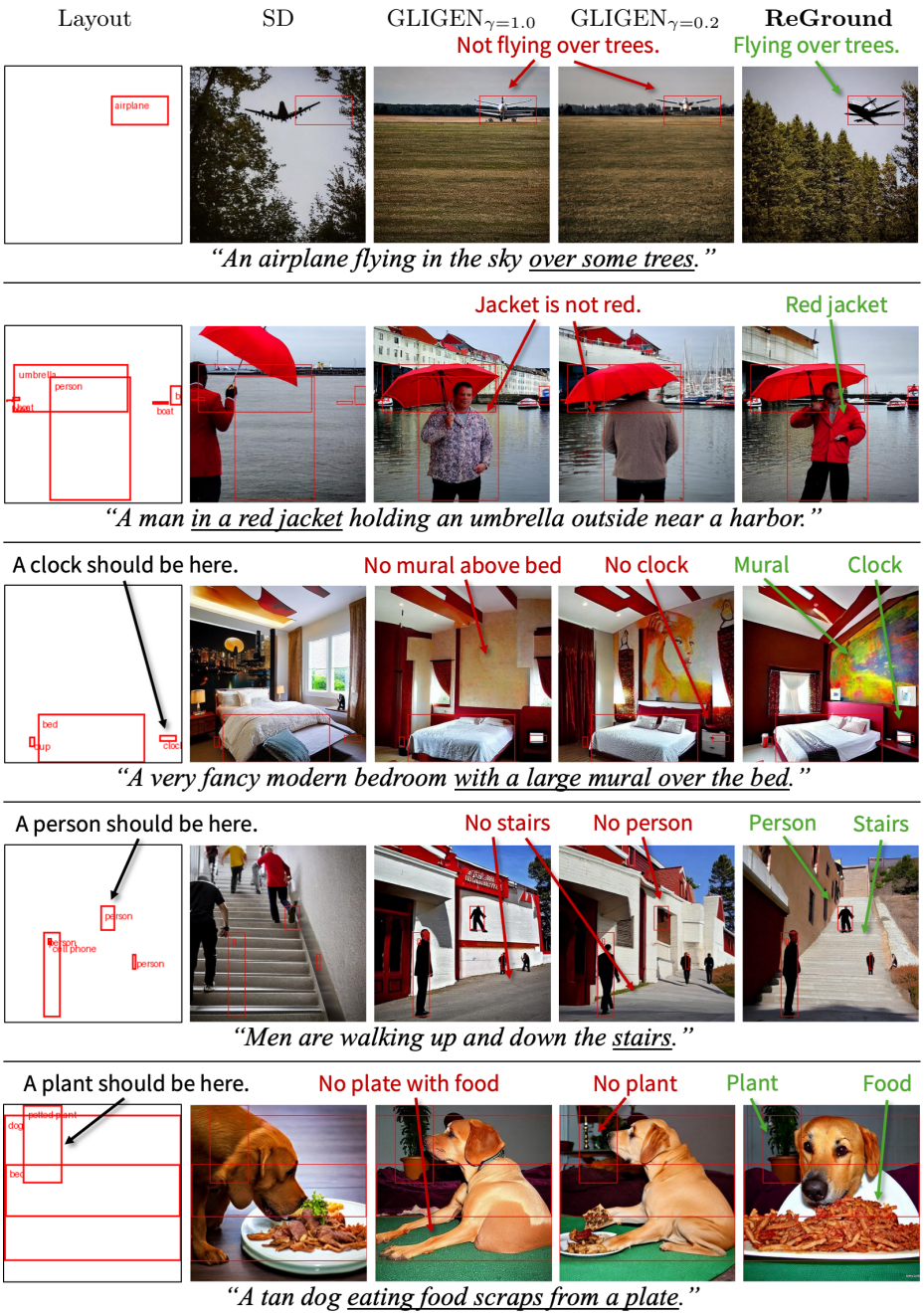
"The men are trying to fly their kite by the water."

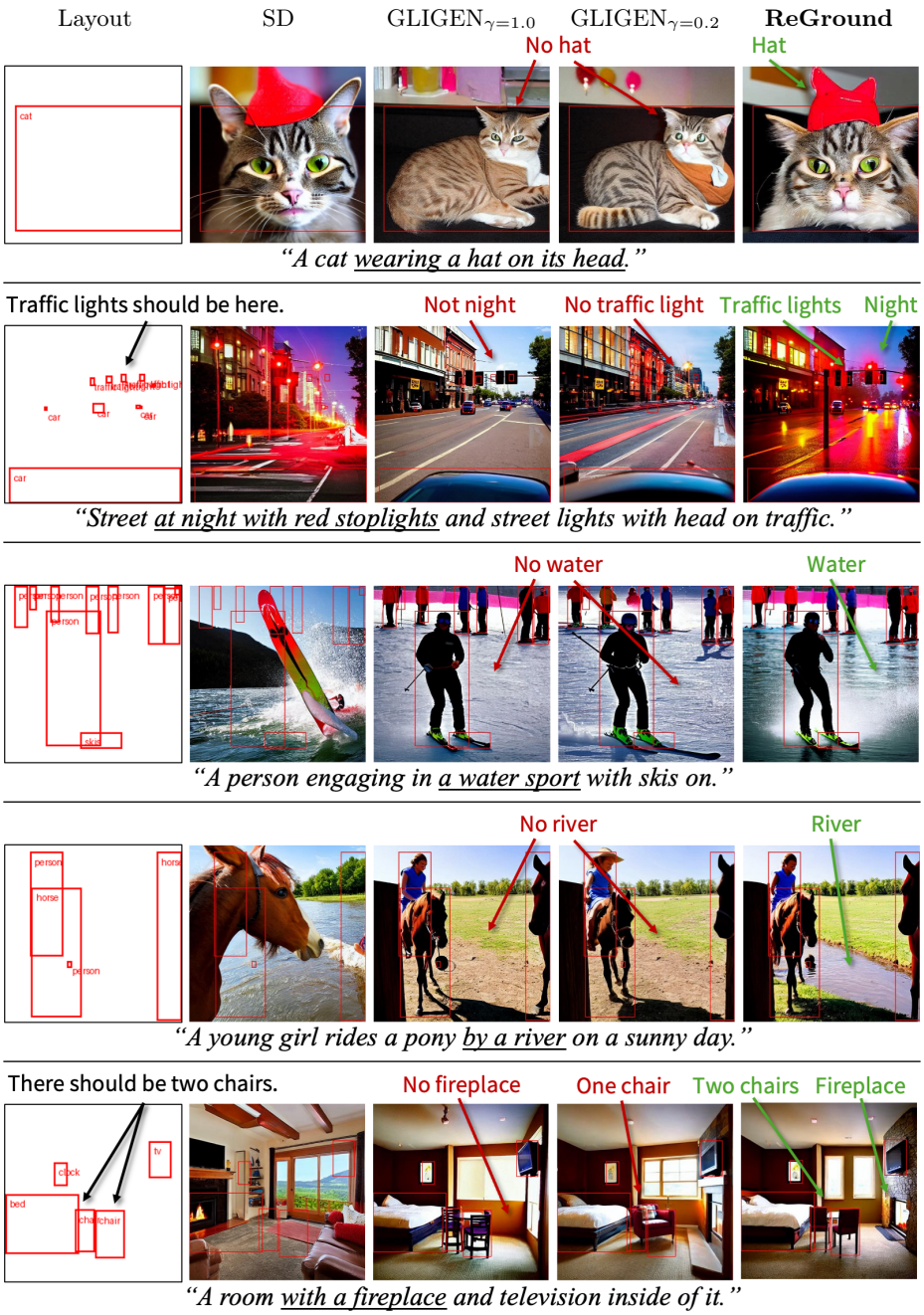


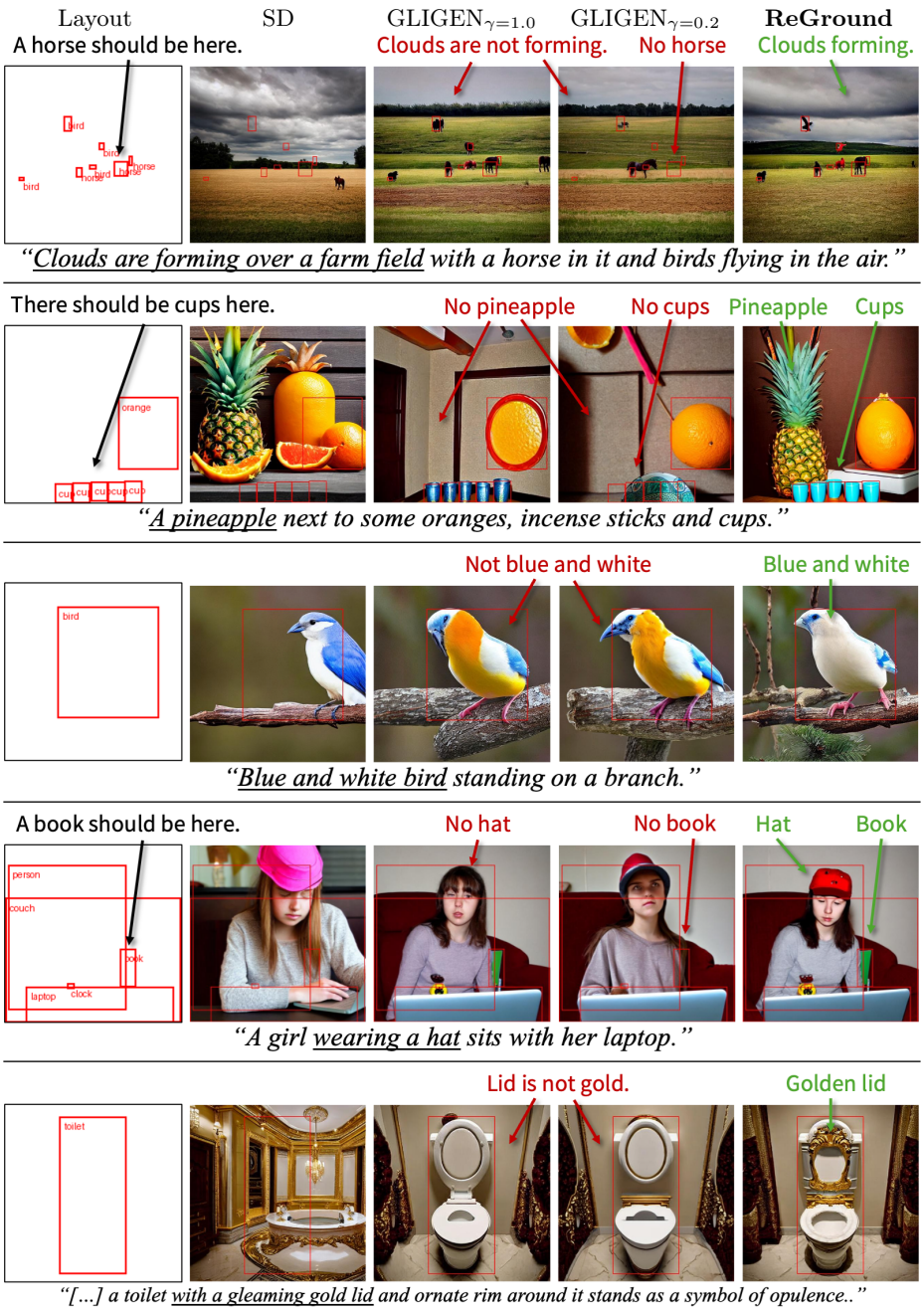
"A dog sitting in an orange chair looking at someone."

"This is a picture of an interesting mural on a truck."

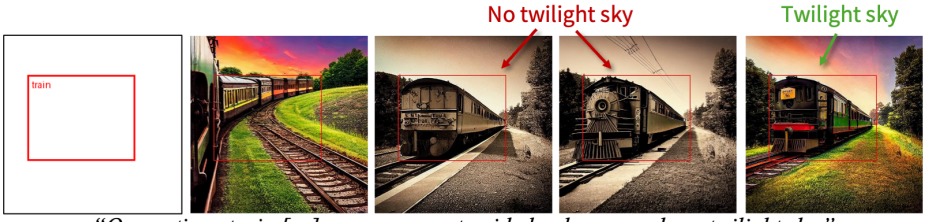

Layout
SD
GLIGEN$_{\gamma=1.0}$
GLIGEN$_{\gamma=0.2}$
ReGround
A bottle should be here.
No water
No bottle
Bottle
Water
refrigerator
truck
chair
bottle
"A refrigerator on a patio flooded with water, beverage cans and a beer bottle."
Not jumping
Jumping over
person
horse
"A woman riding a horse jumps over an obstacle."
A cat should be here.
Not a black cat
Cat out of box
Black cat
motorcycle
cat
"A black cat meanders up the walkway in front of a motorcycle in front of a house."
Traffic lights should be here.
Not a forest
No traffic light
Green forest
traffic light
"A road in the middle of a beautiful green forest."
Train is not green and black.
Green and black train
train
"A person standing next to a green and black train."


Layout
SD
GLIGEN$_{\gamma=1.0}$
GLIGEN$_{\gamma=0.2}$
ReGround
Not flying over trees.
Flying over trees.
airplane
"An airplane flying in the sky over some trees."
Jacket is not red.
Red jacket
umbrella
person
boat
"A man in a red jacket holding an umbrella outside near a harbor."
A clock should be here.
No mural above bed
No clock
Mural
Clock
bed
clock
"A very fancy modern bedroom with a large mural over the bed."
A person should be here.
No stairs
No person
Person
Stairs
person
cell phone
"Men are walking up and down the stairs."
A plant should be here.
No plate with food
No plant
Plant
Food
potted plant
dog
bed
"A tan dog eating food scraps from a plate."


Layout
SD
GLIGEN$_{\gamma=1.0}$
GLIGEN$_{\gamma=0.2}$
ReGround
No hat
Hat
cat
"A cat wearing a hat on its head."
Traffic lights should be here.
Not night
No traffic light
Traffic lights
Night
car
"Street at night with red stoplights and street lights with head on traffic."
No water
Water
person
ski
"A person engaging in a water sport with skis on."
No river
River
person
horse
"A young girl rides a pony by a river on a sunny day."
There should be two chairs.
No fireplace
One chair
Two chairs
Fireplace
bed
chair
"A room with a fireplace and television inside of it."

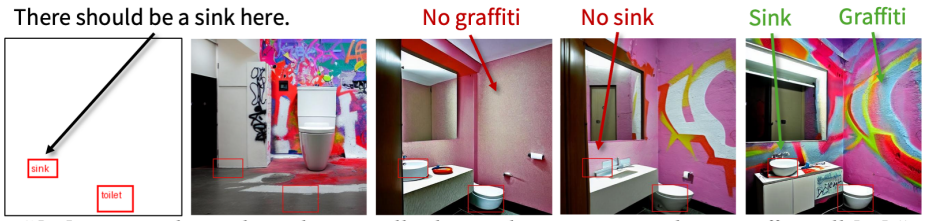

Layout
SD
GLIGEN$_{\gamma=1.0}$
GLIGEN$_{\gamma=0.2}$
ReGround
A horse should be here.
Clouds are not forming.
No horse
Clouds forming.
"Clouds are forming over a farm field with a horse in it and birds flying in the air."
There should be cups here.
No pineapple
No cups
Pineapple
Cups
"A pineapple next to some oranges, incense sticks and cups."
Not blue and white
Blue and white
"Blue and white bird standing on a branch."
A book should be here.
No hat
No book
Hat
Book
"A girl wearing a hat sits with her laptop."
Lid is not gold.
Golden lid
"[...] a toilet with a gleaming gold lid and ornate rim around it stands as a symbol of opulence.."

"Two antique clocks are in the photo, set against a surreal backdrop of a swirling, starlit sky."

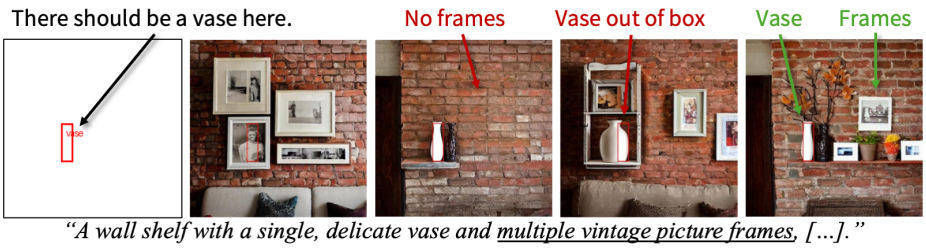


"One antique train [...] a serene countryside landscape under a twilight sky."

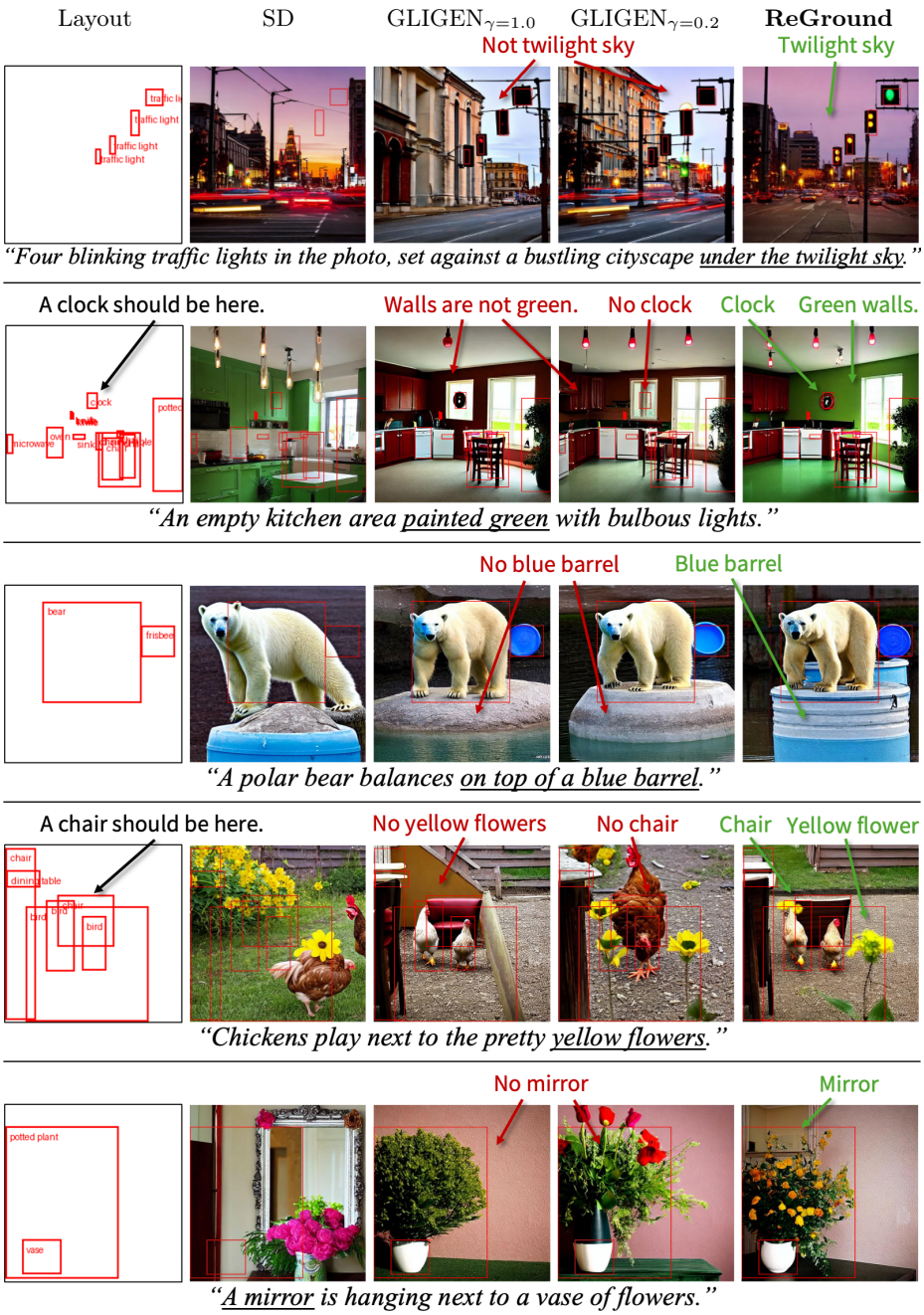
"[...] orange carrot together with one majestic elephant, [...] a surreal, rainbow-colored savannah."

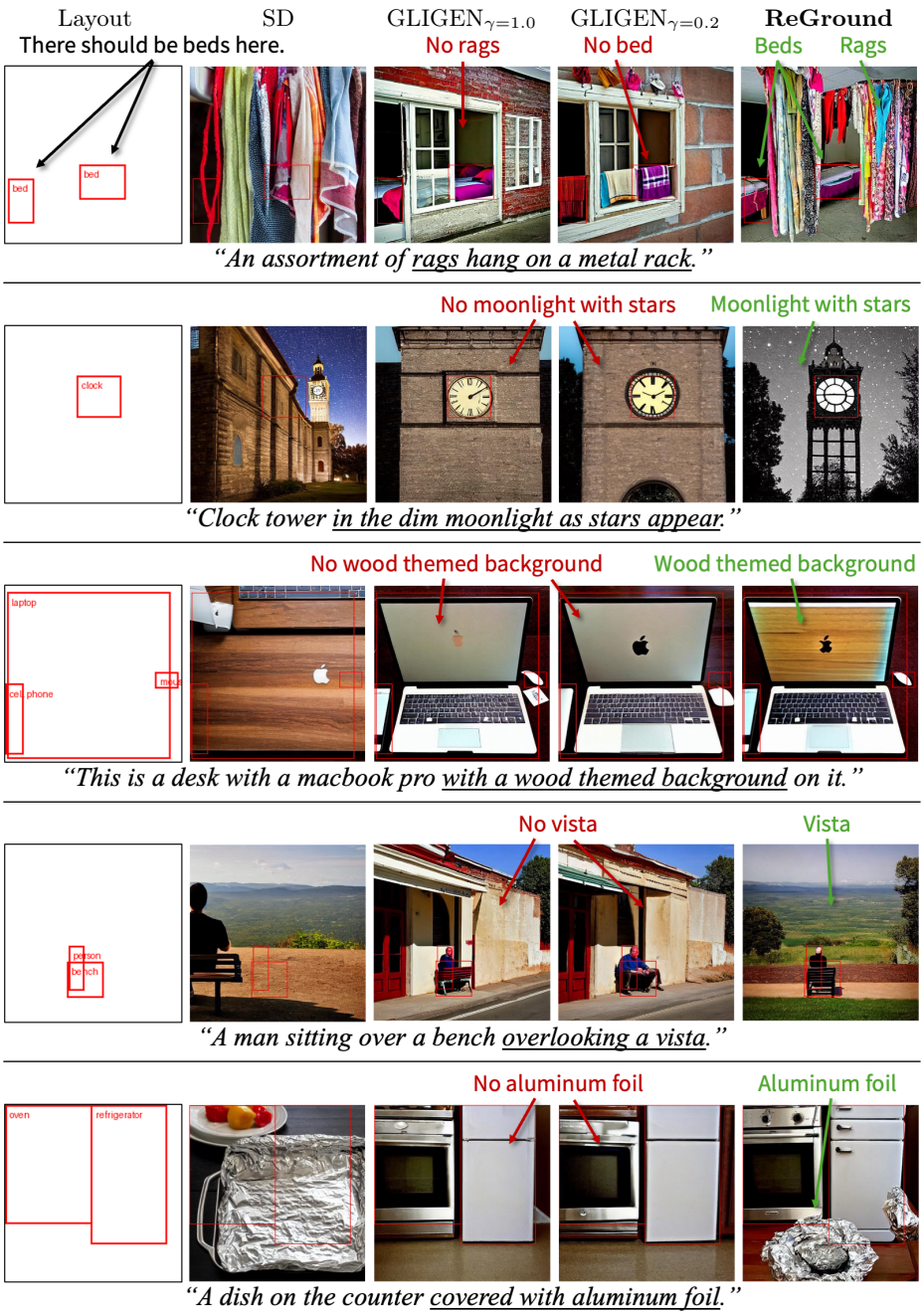


"[...] pristine white toilet with an equally clean sink, set against a vibrant graffiti wall [...]."

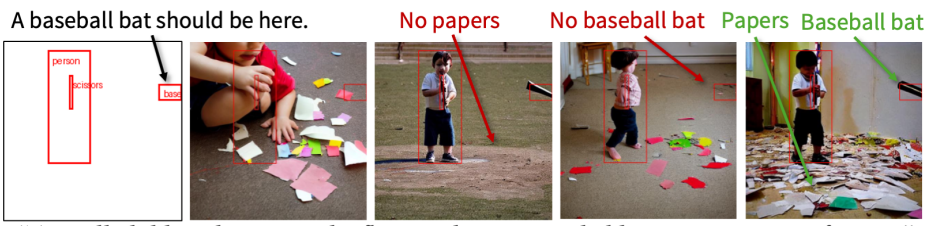


"A wall shelf with a single, delicate vase and multiple vintage picture frames, [...]."


Layout
SD
GLIGEN$_{\gamma=1.0}$
GLIGEN$_{\gamma=0.2}$
ReGround
Not twilight sky
Twilight sky
traffic light
traffic light
traffic light
traffic light
"Four blinking traffic lights in the photo, set against a bustling cityscape under the twilight sky."
A clock should be here.
Walls are not green.
No clock
Clock
Green walls.
clock
potted
knife
oven
microwave
sink
chair
table
"An empty kitchen area painted green with bulbous lights."
No blue barrel
Blue barrel
bear
frisbee
"A polar bear balances on top of a blue barrel."
A chair should be here.
No yellow flowers
No chair
Chair
Yellow flower
chair
dining table
bird
bird
bird
"Chickens play next to the pretty yellow flowers."
No mirror
Mirror
potted plant
vase
"A mirror is hanging next to a vase of flowers."

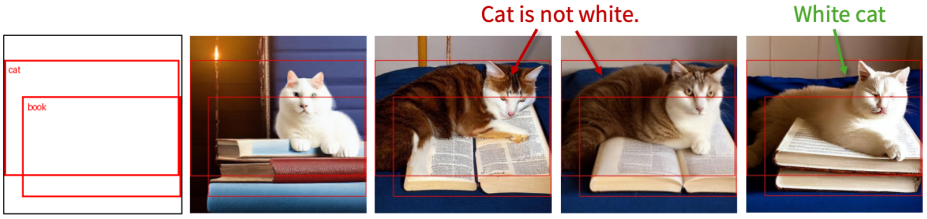

Layout
SD
GLIGEN$_{\gamma=1.0}$
GLIGEN$_{\gamma=0.2}$
ReGround
There should be beds here.
No rags
No bed
Beds
Rags
bed
bed
"An assortment of rags hang on a metal rack."
No moonlight with stars
Moonlight with stars
clock
"Clock tower in the dim moonlight as stars appear."
No wood themed background
Wood themed background
laptop
cell phone
"This is a desk with a macbook pro with a wood themed background on it."
No vista
Vista
person
bench
"A man sitting over a bench overlooking a vista."
No aluminum foil
Aluminum foil
oven
refrigerator
"A dish on the counter covered with aluminum foil."

"A desk in a classroom with a blackboard behind it."



"A small child is playing on the floor and is surrounded by torn up pieces of paper."

"A stainless steel oven to the left of a sleek refrigerator in a modern kitchen [...] a view of the city skyline."



"[...] a fluffy, white cat is peacefully resting its head on an ancient, leather-bound book on a plush, blue bed."

References

1. Feng, W., Zhu, W., Fu, T.j., Jampani, V., Akula, A., He, X., Basu, S., Wang, X.E., Wang, W.Y.: LayoutGPT: Compositional visual planning and generation with large language models. In: NeurIPS (2023) 1
2. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: CLIPScore: a reference-free evaluation metric for image captioning. In: EMNLP (2021) 2
3. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: GLIGEN: Open-set grounded text-to-image generation. In: CVPR (2023) 1, 2, 3, 5
4. Lian, L., Li, B., Yala, A., Darrell, T.: Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. arXiv preprint arXiv:2305.13655 (2023) 3
5. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014) 1
6. OpenAI: Chatgpt, https://chat.openai.com/ 1
7. Phung, Q., Ge, S., Huang, J.B.: Grounded text-to-image synthesis with attention refocusing. arXiv preprint arXiv:2306.05427 (2023) 3, 4
8. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022) 1, 2, 5
9. Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M.: YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: CVPR (2023) 2
10. Xie, J., Li, Y., Huang, Y., Liu, H., Zhang, W., Zheng, Y., Shou, M.Z.: Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In: ICCV (2023) 3, 4
11. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018) 1