



REVISION: Rendering Tools Enable Spatial Fidelity in Vision-Language Models: Supplementary Materials

Agneet Chatterjee ^{*1}, Yiran Luo ^{*1},
Tejas Gokhale², Yezhou Yang¹, and Chitta Baral¹

¹ Arizona State University

² University of Maryland, Baltimore County

In this supplementary material, we present additional results on ControlNet and GPT-4 Guided Coordinate Generation results. We also present illustrative samples, covering successful and failure image generation as well as results on the human evaluation experiments. Lastly, we present asset samples from the REVISION pipeline along with outputs from the Position Diversifier.

1 Additional Quantitative Results

The ControlNet-based results are presented in Table [1](#). We achieve the best trade-off between IS and VISOR for ControlNet and compared to Stable Diffusion, we achieve a higher VISOR₄ score, indicating correctness over multiple trials. Thus, we quantify that REVISION generated images have enough low-level information to faithfully represent spatial orientations.

2 GPT-4 Guided Co-ordinate Generation

We also experiment with generating flexible coordinates for the objects using GPT-4. We first feed GPT-4 a designed in-context prompt that includes specific example coordinates for each possible spatial relation. We then feed in an input prompt of two objects and one spatial relation in order to obtain the two sets of coordinates for placing the mentioned objects. Table [2](#) shows results of performing conditioning using GPT-4 as the alternative Coordinate Generator. Compared to our baseline results in, we notice an average of 10-point drop in performance in both Object Accuracy and VISOR_{uncond} scores. These patterns develop as a result of GPT-4’s propensity to generate co-ordinates which places the two objects in close proximity, leading to images where the objects are indistinguishable. Hence, T2I models tend to ignore either object, which correspondingly lead to lower VISOR scores.

* Equal contribution. Correspondence to agneet@asu.edu

Table 1: ControlNet + REVISION results on VISOR. As indicated by the high VISOR₄ score, we consistently generate images which are spatially correct.

Background	IS	OA (%)	VISOR (%)					
			uncond	cond	1	2	3	4
White	18.82	56.88	55.48	97.54	78.82	62.93	48.58	31.59
Indoor	14.75	59.64	58.08	97.39	81.38	66.35	51.27	33.33
Outdoor	<u>16.20</u>	56.54	<u>56.22</u>	99.45	75.97	<u>62.99</u>	<u>50.57</u>	35.37

Table 2: VISOR Results on using GPT-4 as the Coordinate Generator. The drop in performance is attributed to the proclivity of GPT-4 to place both the objects too close to each other in the coordinate space.

Method	OA	VISOR	
		uncond	cond
SD 1.4 + REVISION	43.88	<u>41.18</u>	93.85
SD 1.5 + REVISION	44.35	41.64	<u>93.89</u>
SD 2.1 + REVISION	39.03	36.86	94.43

3 Object-Wise Spatial Accuracy Analysis

In Figure 1 we show the success rate of correctly generating objects from MS-COCO using REVISION-based guidance. On average, there is a 61% likelihood that an MS-COCO object is accurately positioned in the output image.

4 Illustrative Results for Human Evaluation Experiments

4.1 Prompts of Multiple Objects and Relationships

We present illustrative results in Figure 2, where we show that REVISION extends accurate image generation when a prompt includes multiple objects and spatial relationships.

4.2 Out-of-Distribution Objects (OOD)

For experiments involving OOD objects, we swap one OOD object in the input prompt with its corresponding MS-COCO substitute object. We present the corresponding substitutes in Table 3 and show illustrative examples in Figure 3.

5 Additional Illustrations

Next, we demonstrate additional illustrations of images generated using images from REVISION as additional guidance; Figure 4 shows successfully generated images with REVISION while Figure 5 presents failure scenarios.

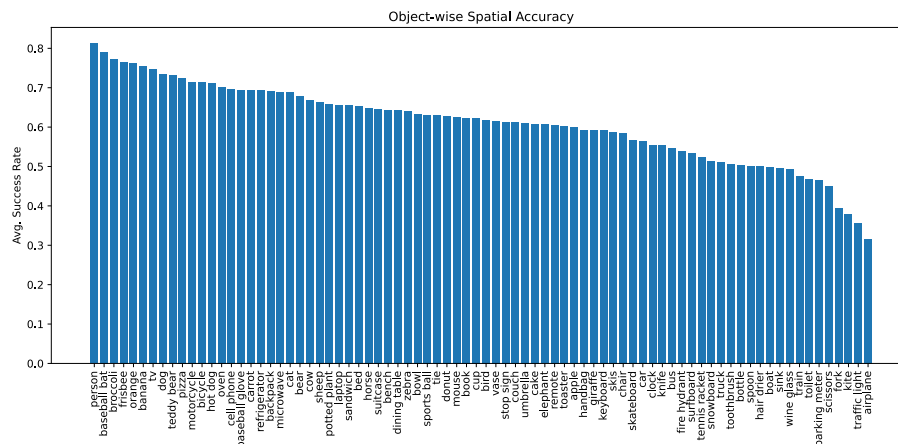


Fig. 1: Average Success Rate of each MS-COCO object in REVISION, being spatially correct according to the input prompt in the generated image. We report results using the white background with SD v1.5.



Fig. 2: Illustrative example of leveraging REVISION to generate spatially correct images with 3 objects and 2 relationships.

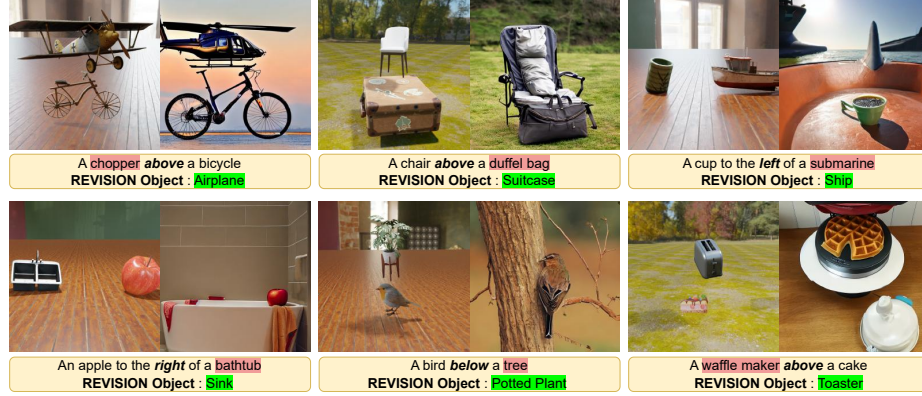


Fig. 3: Illustrative example of leveraging REVISION to generate images with objects not in our asset library. For each pair of image, left is the reference image from REVISION, and right is the generated image. Objects in **green** are from our asset library, while objects in **red** are OOD objects.

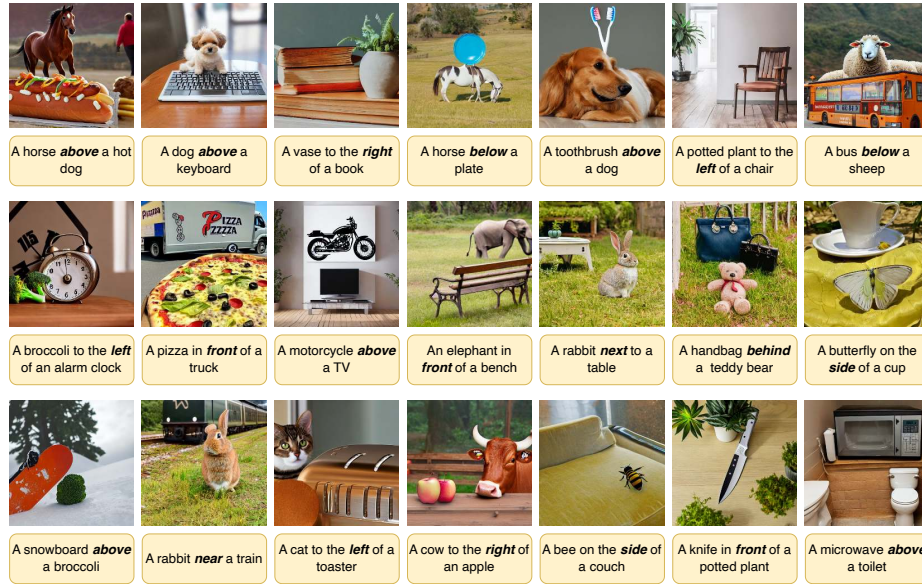


Fig. 4: Correctly generated images from T2I models by leveraging images from REVISION as additional guidance.

Table 3: Substitute OOD object nouns for the original 80 MS-COCO objects used in REVISION.

MS-COCO Object	OOD Object	MS-COCO Object	OOD Object
airplane	helicopter	kite	flag
apple	pear	knife	sword
backpack	purse	laptop	tablet
banana	mango	microwave	toaster oven
baseball bat	walking stick	motorcycle	tractor
baseball glove	boxing glove	mouse	webcam
bear	monkey	orange	papaya
bed	table	oven	dishwasher
bench	sofa	parking meter	phone booth
bicycle	scooter	person	mannequin
bird	butterfly	pizza	burger
boat	submarine	potted plant	tree
book	magazine	refrigerator	cabinet
bottle	lunchbox	remote	game controller
bowl	plate	sandwich	salad
broccoli	cauliflower	scissors	pliers
bus	tram	sheep	goat
cake	pie	sink	bathtub
car	ambulance	skateboard	roller skates
carrot	sweet potato	skis	hockey stick
cat	rabbit	snowboard	sled
cell phone	landline phone	spoon	straw
chair	barstool	sports ball	bowling ball
clock	wall calendar	stop sign	parking sign
couch	cushion	suitcase	duffel bag
cow	panda	surfboard	kayak
cup	tumbler	teddy bear	doll
dining table	dressing table	tennis racket	badminton racket
dog	fox	tie	bowtie
donut	pudding	toaster	waffle maker
elephant	lion	toilet	shower
fire hydrant	mailbox	toothbrush	comb
fork	chopsticks	traffic light	streetlight
frisbee	basketball	train	roller coaster
giraffe	camel	truck	crane
hair drier	hairbrush	tv	computer monitor
handbag	cardboard box	umbrella	tent
horse	donkey	vase	pitcher
hot dog	burrito	wine glass	glass jar
keyboard	piano	zebra	llama

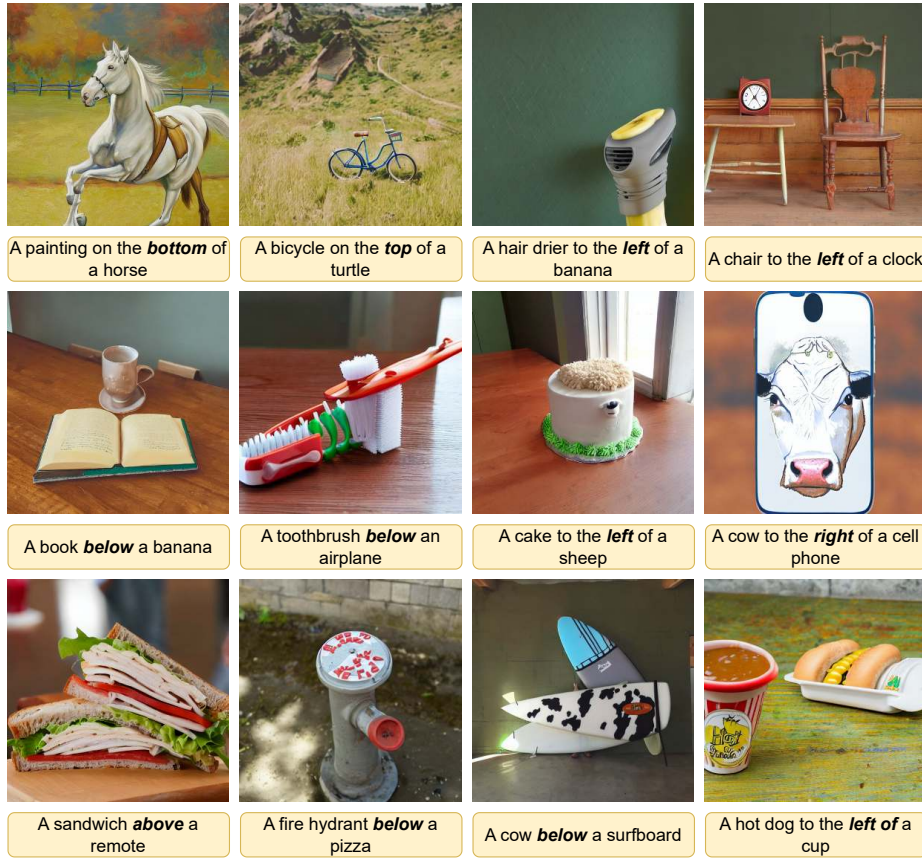


Fig. 5: Images generated from T2I models by leveraging REVISION, which either do not contain correct objects or are spatially incorrect.

6 REVISION Assets and Illustrations

Figure 6 and 7 illustrate instances of MS-COCO 3D assets present in REVISION, organized by subcategory. Figure 8 presents the non MS-COCO objects and their corresponding class labels. We present results from our Position Diversifier module in Figure 9. For a given set of assets and spatial relationship, we generate multiple distinct instances. In Figure 10 and 11, we present images from REVISION which depict the *near* and *depth* spatial relationships, respectively.

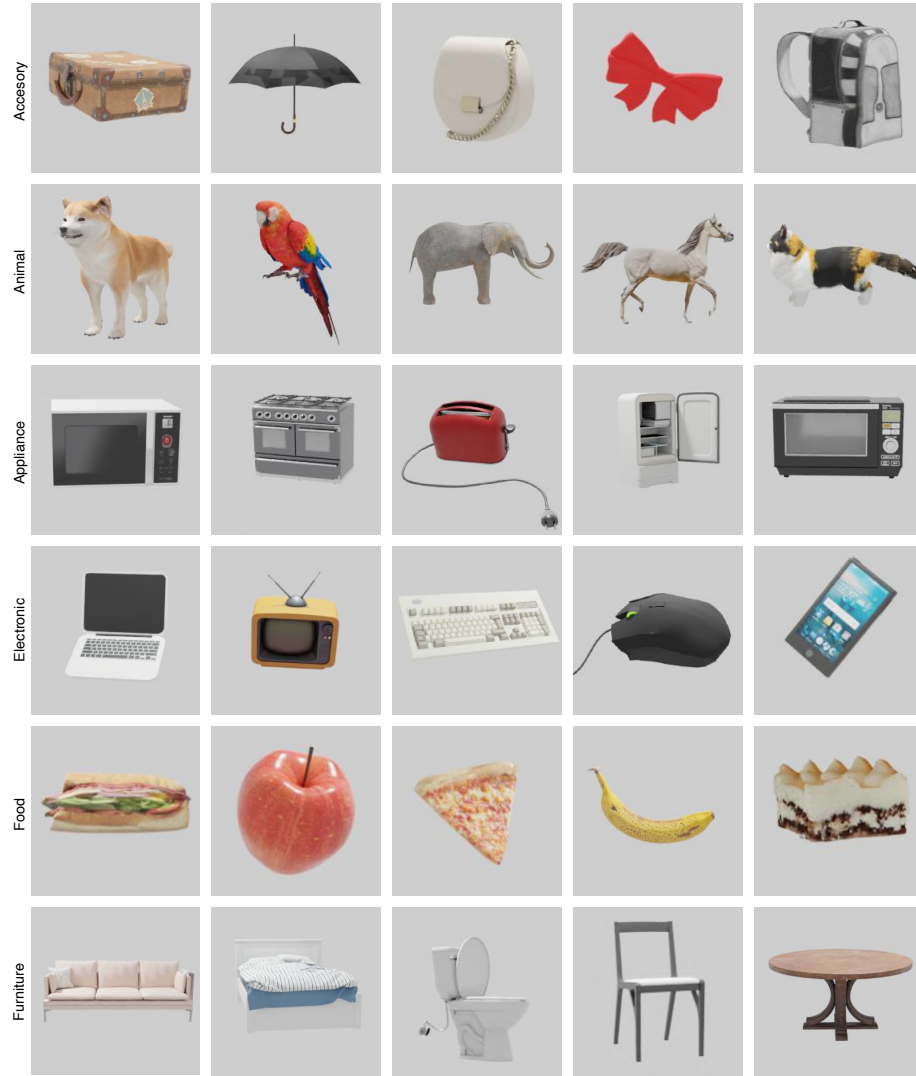


Fig. 6: Example 3D models of MSCOCO objects featured in REVISION’s Asset Library.

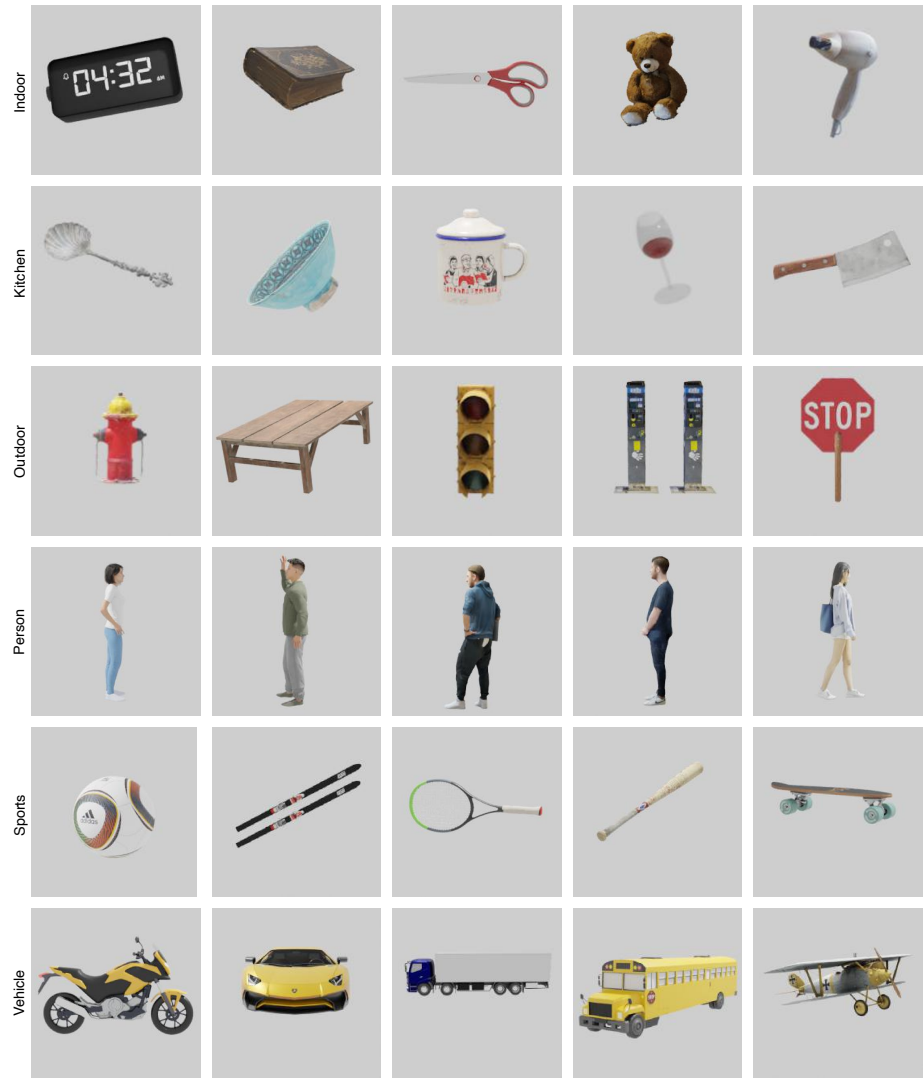


Fig. 7: Example 3D models of MSCOCO objects featured in REVISION’s Asset Li-
brary, continued.

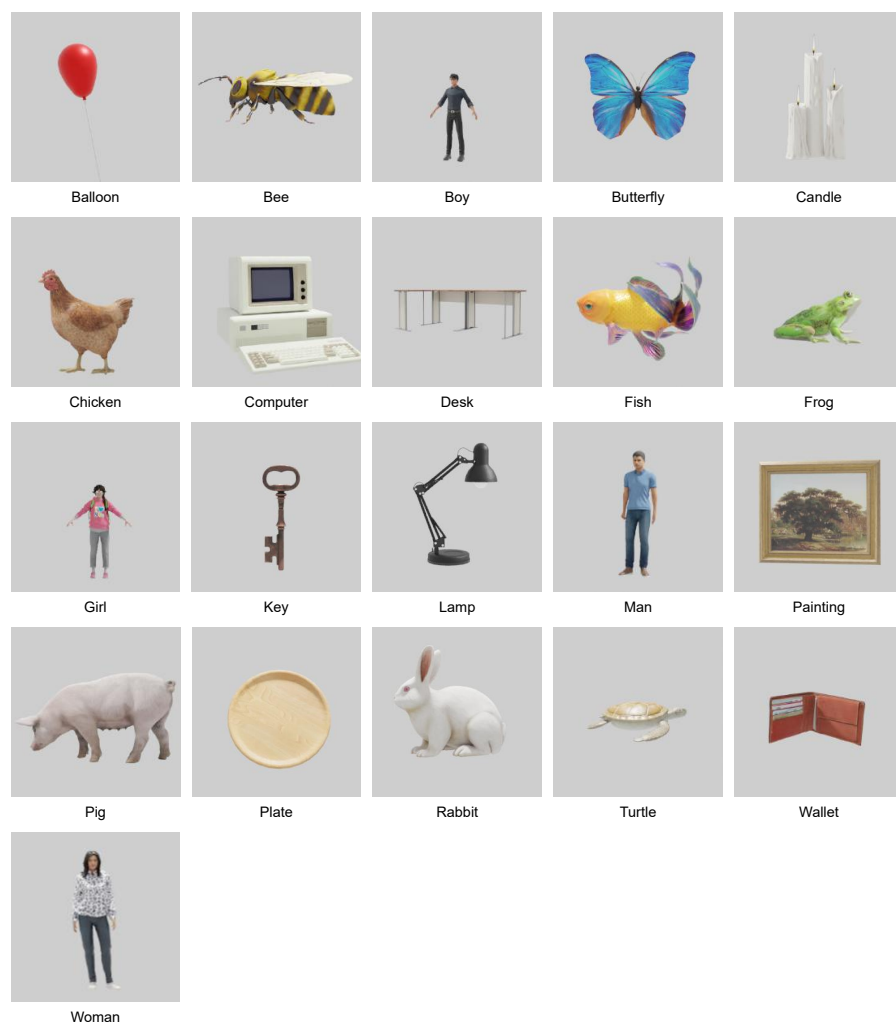


Fig. 8: Example 3D models of Non-MSCOCO objects featured in REVISION's Asset Library.

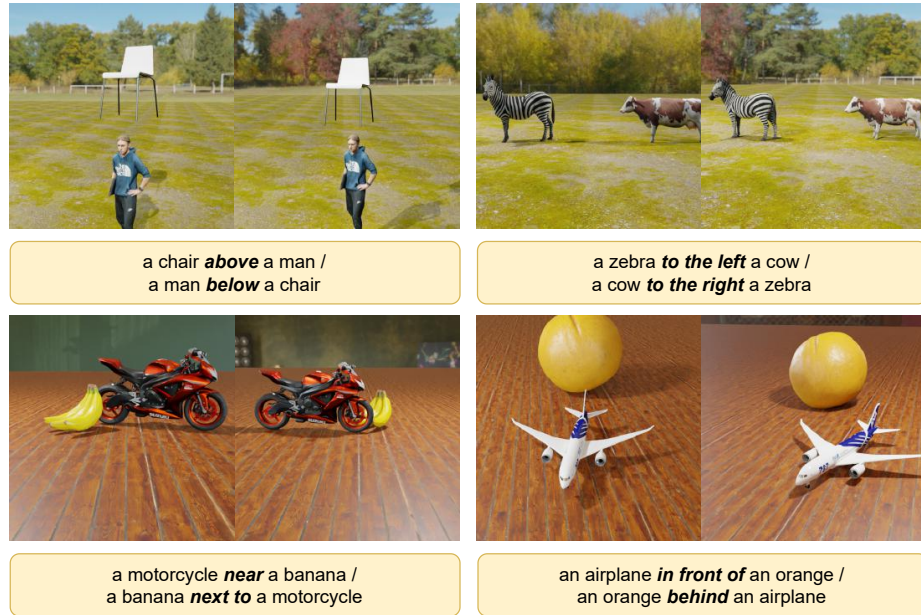


Fig. 9: Diversified scenes achieved by the Position Diversifier of REVISION, in all categories of spatial relationships.



Fig. 10: Example REVISION outputs in *near* relationship, featuring two object assets within close proximity or touching each other.

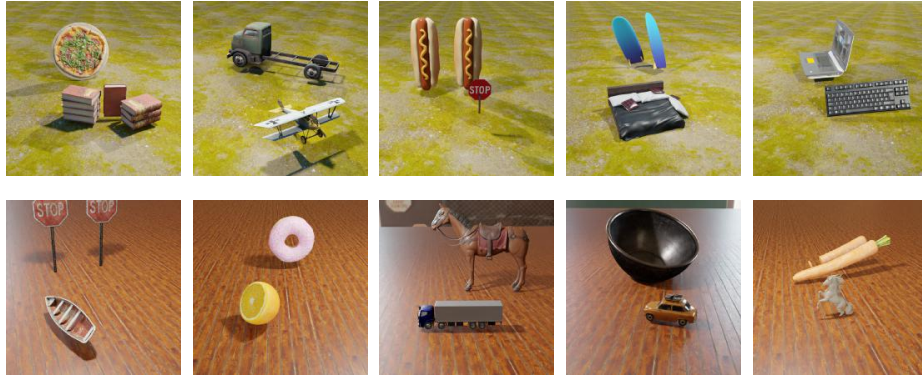


Fig. 11: Example REVISION outputs in *depth* relationship, featuring two object assets in front of/behind one another. The angle of the camera is also relatively elevated to strengthen the depth perspective.