

A Detailed scoring criteria for each sub-domain.

In Tab.1, we show the demand for spatial difference for different subdomains.

A.1 Spatial Pose

For the two subdomains Camera Pose and Object Pose, there is only one object in the image, and the criterion for judging is the orientation of the object, so there is only one difference that needs to be introduced, orientation difference. For Complex Pose, which involves multiple objects, it is necessary to maintain the side-by-side and same orientation relationship of these objects in the text-to-image generation task; while in the image editing task, it is necessary to maintain the general spatial relationship between the objects before and after editing without any major changes. Therefore, in addition to orientation difference, it is necessary to introduce relation difference.

A.2 Spatial Relation

All three subdomains of this dimension encompass multiple objects. Both orientation and relation differences are required in the scoring stage.

A.3 Spatial Measurement

In this domain, Distance Difference is necessary in order to measure the length/width/height/volume of an object. Besides, Orientation difference is required for the subdomain of Object Size, because we consider the measure parallel to the front-back direction of an object as length, the measure parallel to the left-right direction of an object as width, and the measure parallel to the top-bottom direction of an object as height. Therefore, for the scoring of Object Size, we need to distinguish the length, width and height of an object by its orientation.

Domain	Sub-domain	Orientation Diff.	Relation Diff.	Distance Diff.
Spatial Pose	Camera Pose	✓	✗	✗
	Object Pose	✓	✗	✗
	Complex Pose	✓	✓	✗
Spatial Relation	Egocentric	✓	✓	✗
	Allocentric	✓	✓	✗
	Intrinsic	✓	✓	✗
Spatial Measurement	Object Size	✓	✗	✓
	Object Distance	✗	✗	✓
	Camera Distance	✗	✗	✓

Table 1: Difference required for each sub-domain

B Impact of Spatial Conditions for More Models

Limitation in camera location understanding. As shown in sub-Fig.a and sub-Fig.b of Figs.1 to 7, most models also suffer from a lack of understanding of distinguishing between side views (e.g. "right/left view") of objects.

Limitation in object location understanding. In addition, the FLUX.1-dev, SD-XL, SD-1.5, and SD-3.5-L show a very clear shortcoming in understanding object orientation (both from the camera pose and the object pose). In most cases they just directly draw the object in front view, no matter what the prompt is.

Limitation in egocentric-allocentric transformation. The sub-Fig.c and sub-Fig.d of Figs.1 to 7 illustrate that the other models are almost exclusively limited to egocentric thinking for object relationships, similar to the GPT-4o.

32 **Limitation in understanding metric measurement.** As with GPT-4o, the other existing models,
 33 almost all of them, are incapable of understanding information from quantitative spatial measurements.
 34 In particular, the 2m in Object Distance and the 4m in Camera Distance are almost rarely generated
 35 by the models, even if prompt tells them to do so, as show in the sub-Fig.e and sub-Fig.f of Figs.1 to
 36 7.

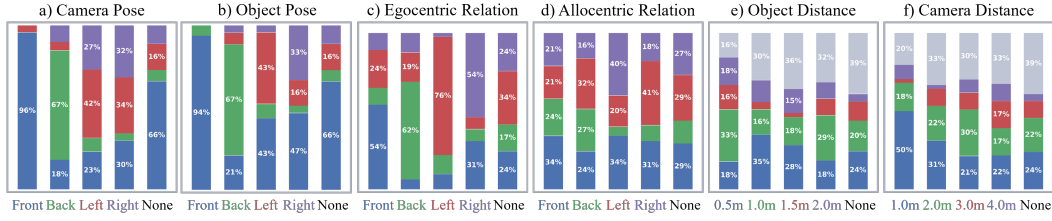


Figure 1: Impact of varying spatial conditions on the spatial states of generated samples from **Gemini-2.5-Pro**.

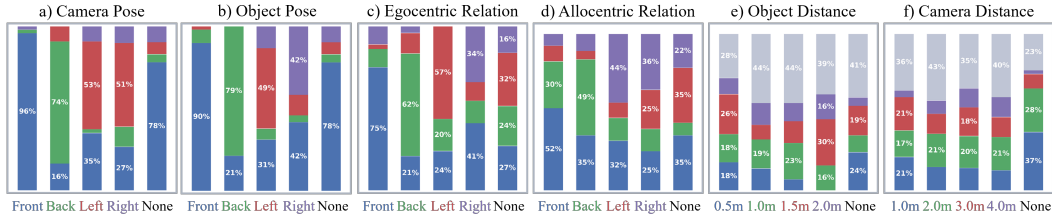


Figure 2: Impact of varying spatial conditions on the spatial states of generated samples from **Seedream-3.0**.

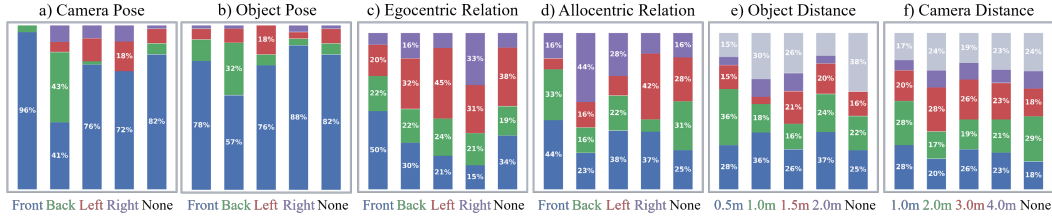


Figure 3: Impact of varying spatial conditions on the spatial states of generated samples from **FLUX.1-dev**.

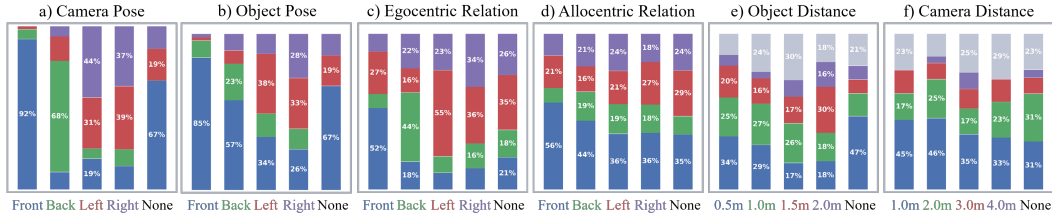


Figure 4: Impact of varying spatial conditions on the spatial states of generated samples from **DALL-E 3**.

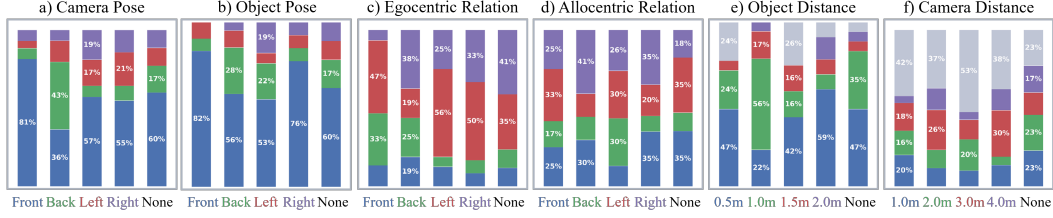


Figure 5: Impact of varying spatial conditions on the spatial states of generated samples from SD-XL.

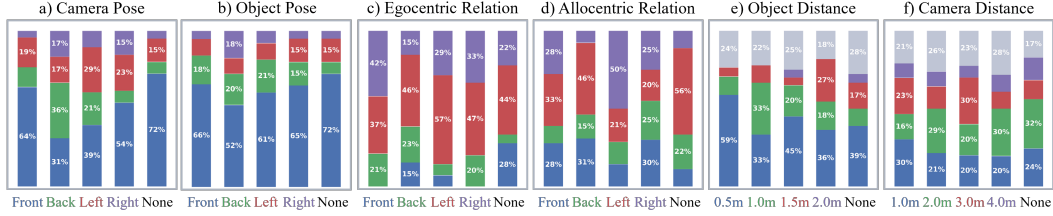


Figure 6: Impact of varying spatial conditions on the spatial states of generated samples from SD-1.5.

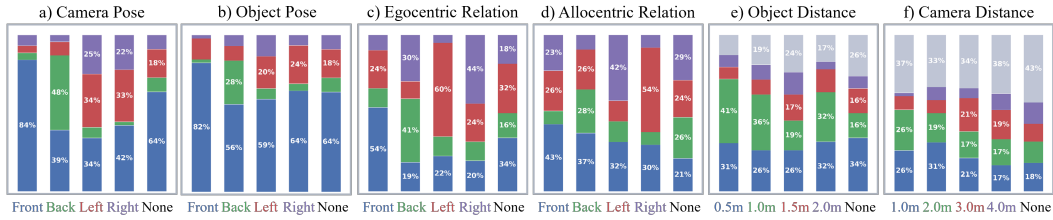


Figure 7: Impact of varying spatial conditions on the spatial states of generated samples from SD-3.5-L.

37 C Visualization of Text-to-image Generation Benchmark

38 In this section, we show the 36 small generation tasks (each subdomain contains 4 tasks) that we
 39 covered in the Text-to-image Generation Benchmark. Each task contains eight images generated by
 40 eight models prompted by the same instruction. The images that match the instruction are labeled
 41 with green boxes, those that do not match are labeled with red models, and those that are partially
 42 correct are labeled with yellow boxes.

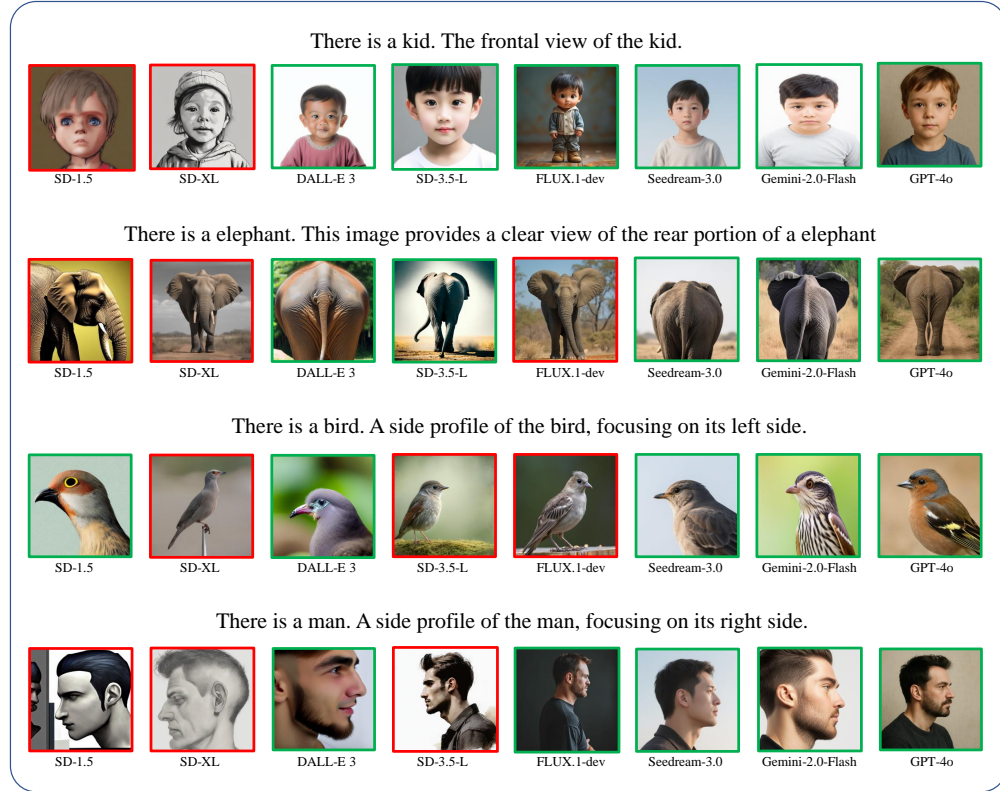


Figure 8: Visualization of Text-to-image Generation Benchmark on the subdomain **Camera Pose**

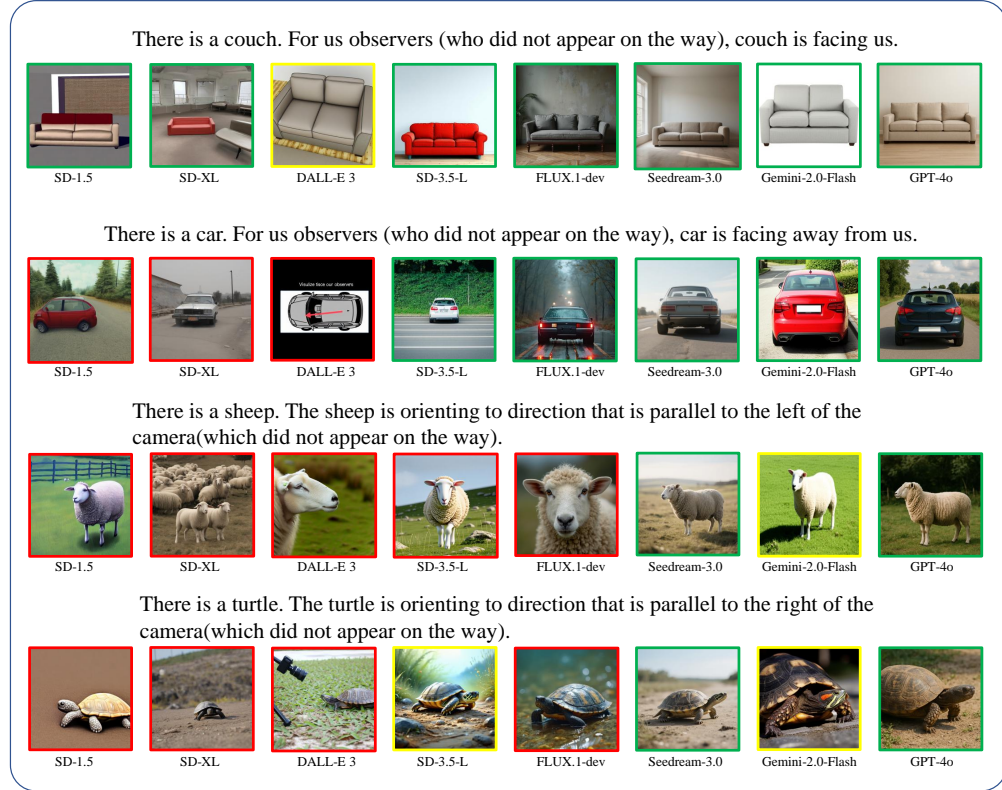


Figure 9: Visualization of Text-to-image Generation Benchmark on the subdomain **Object Pose**

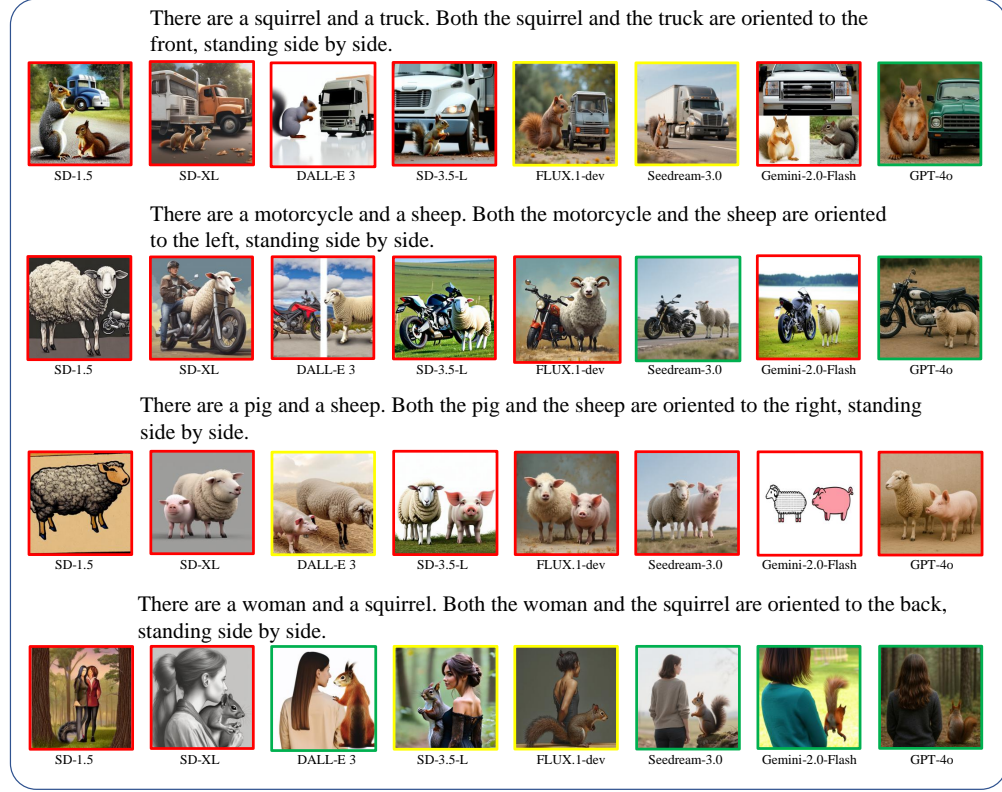


Figure 10: Visualization of Text-to-image Generation Benchmark on the subdomain **Complex Pose**



Figure 11: Visualization of Text-to-image Generation Benchmark on the subdomain **Egocentric Relation**

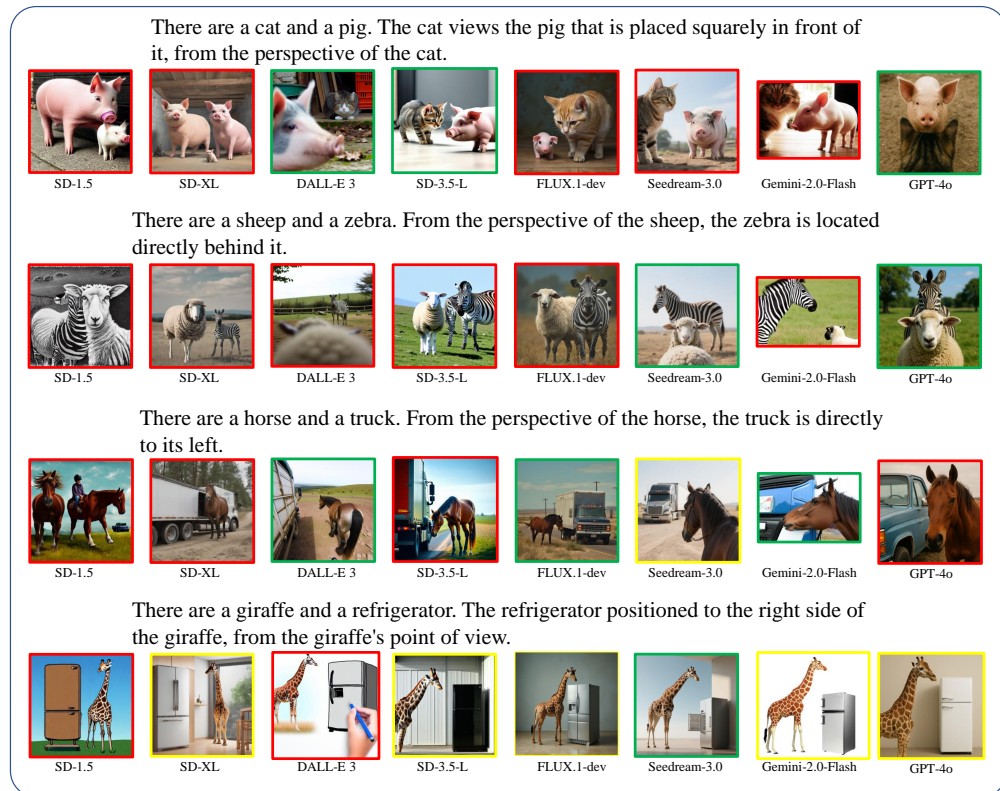


Figure 12: Visualization of Text-to-image Generation Benchmark on the subdomain **Allocentric Relation**

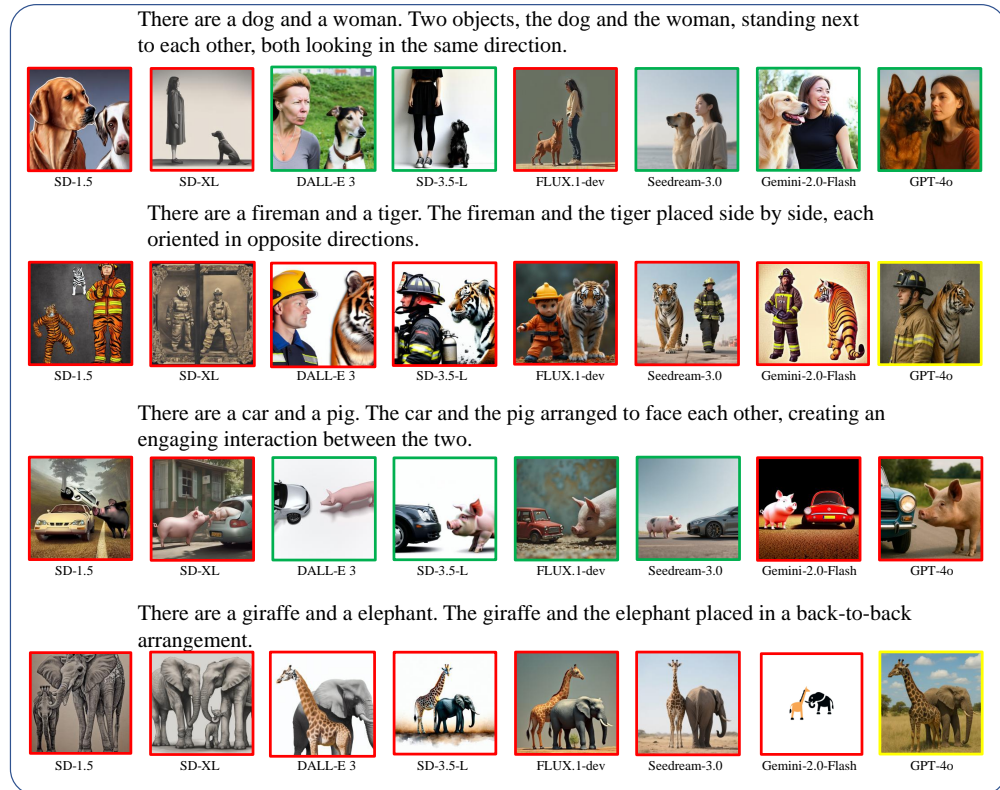


Figure 13: Visualization of Text-to-image Generation Benchmark on the subdomain **Intrinsic Relation**

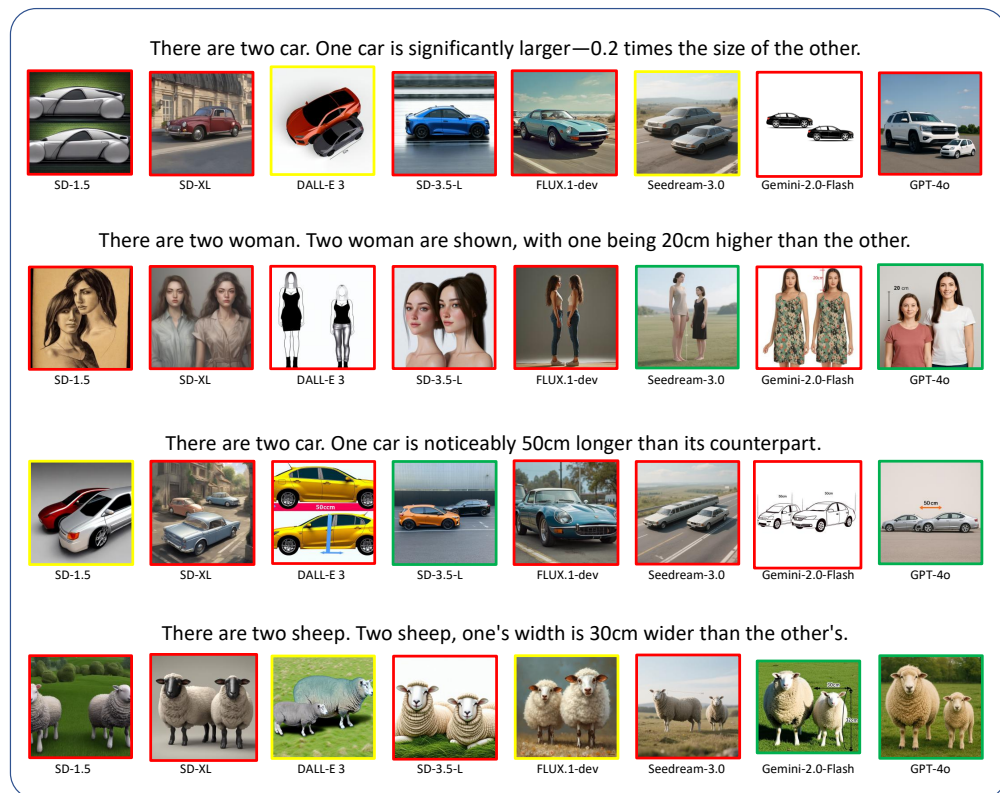


Figure 14: Visualization of Text-to-image Generation Benchmark on the subdomain **Object Size**



Figure 15: Visualization of Text-to-image Generation Benchmark on the subdomain **Object Distance**

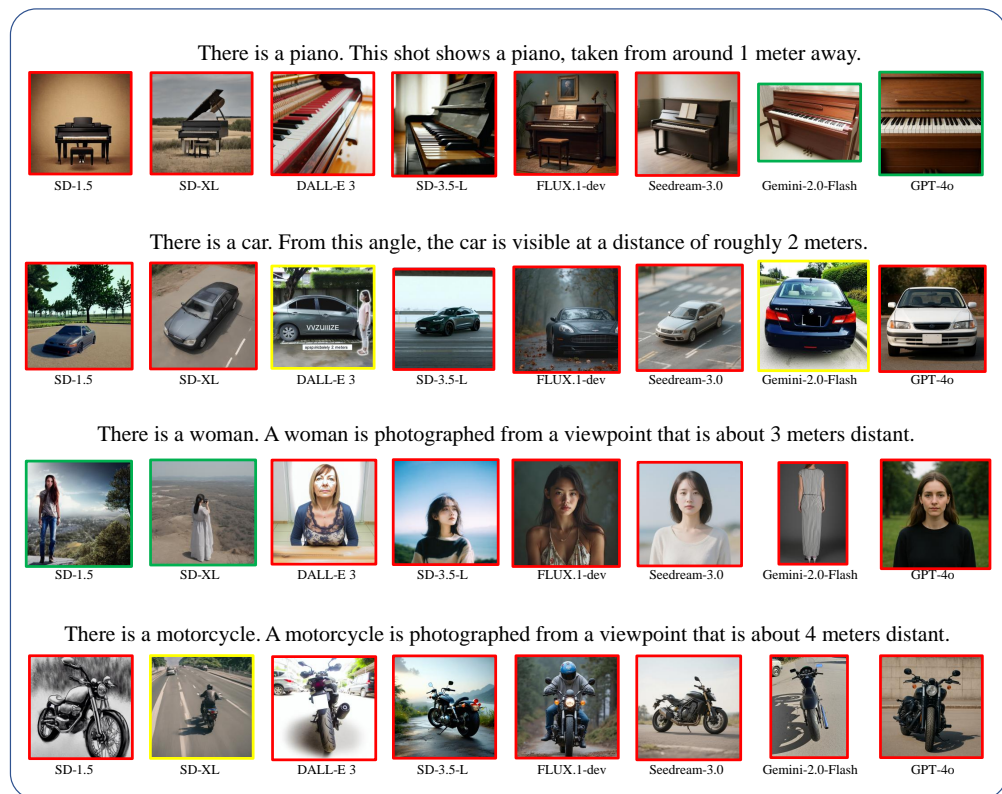


Figure 16: Visualization of Text-to-image Generation Benchmark on the subdomain **Camera Distance**

43 D Visualization of Instruction-based Image Editing Benchmark

44 In this section, we show the 36 small edit tasks (each subdomain contains 4 tasks) that we covered in
 45 the Instruction-based Image Editing Benchmark. Each task contained one original image, and six
 46 images edited by six models from the same text prompt. The images that match the instruction are
 47 labeled with green boxes, those that do not match are labeled with red models, and those that are
 48 partially correct are labeled with yellow boxes.

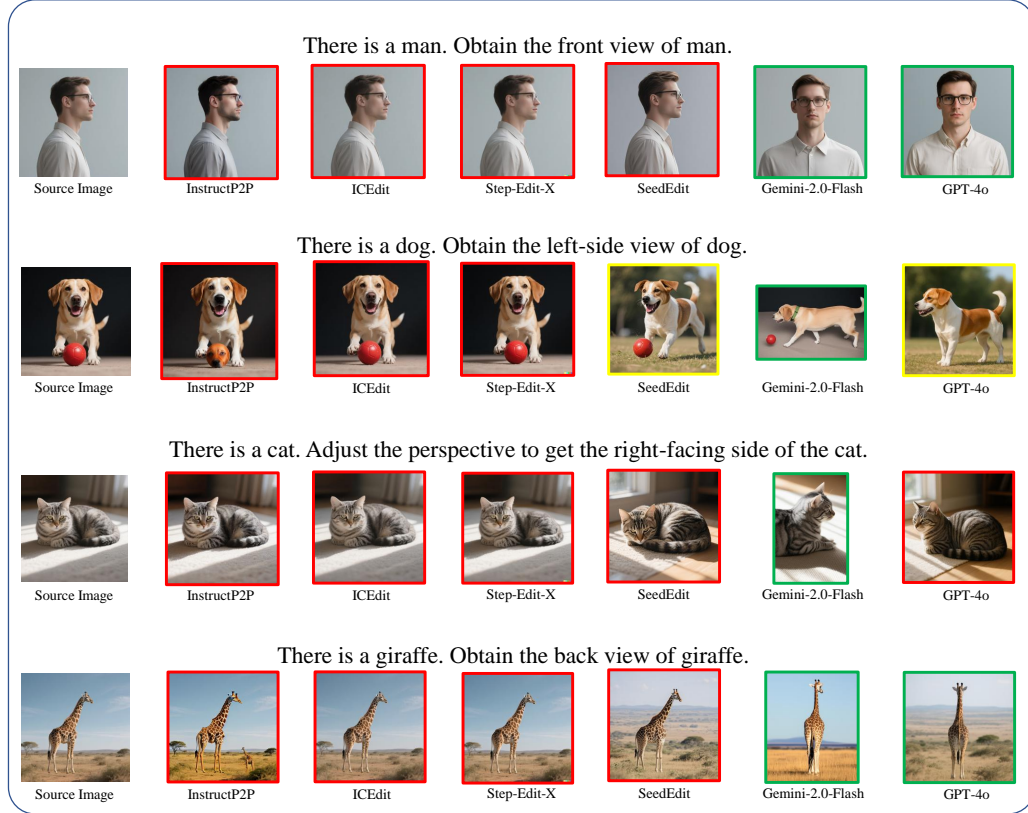


Figure 17: Visualization of Instruction-based Image Editing Benchmark on the subdomain **Camera Pose**

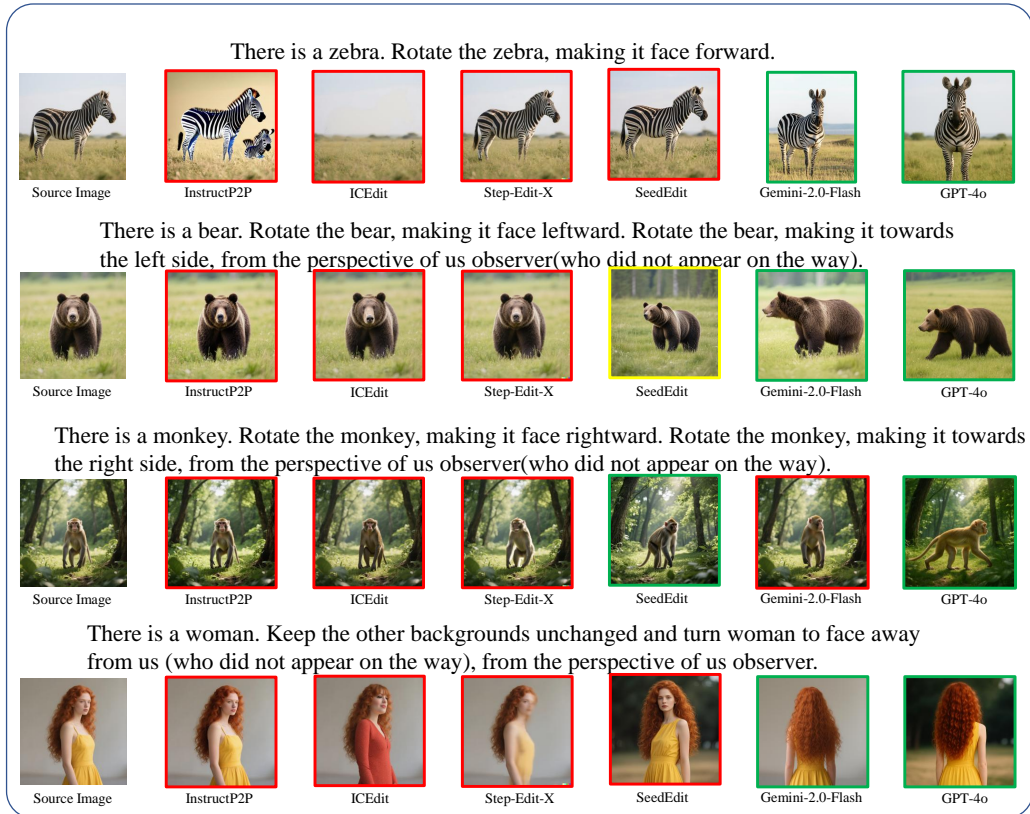


Figure 18: Visualization of Instruction-based Image Editing Benchmark on the subdomain **Object Pose**

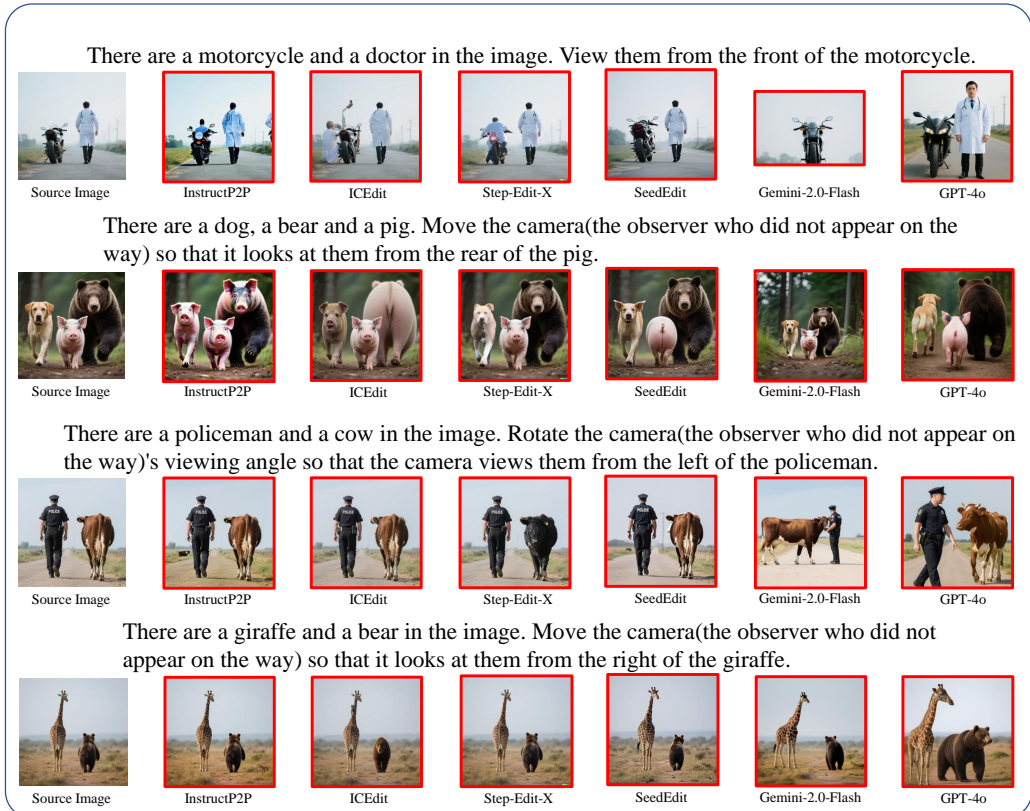


Figure 19: Visualization of Instruction-based Image Editing Benchmark on the subdomain **Complex Pose**

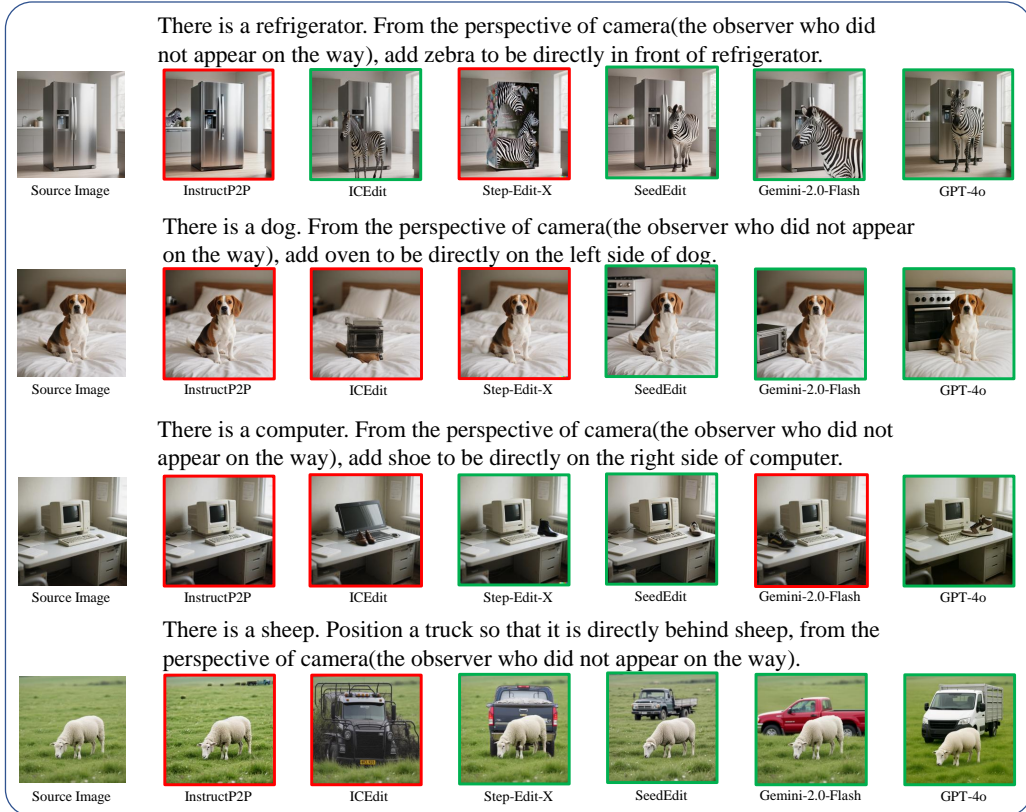


Figure 20: Visualization of Instruction-based Image Editing Benchmark on the subdomain **Egocentric Relation**



Figure 21: Visualization of Instruction-based Image Editing Benchmark on the subdomain **Allocentric Relation**

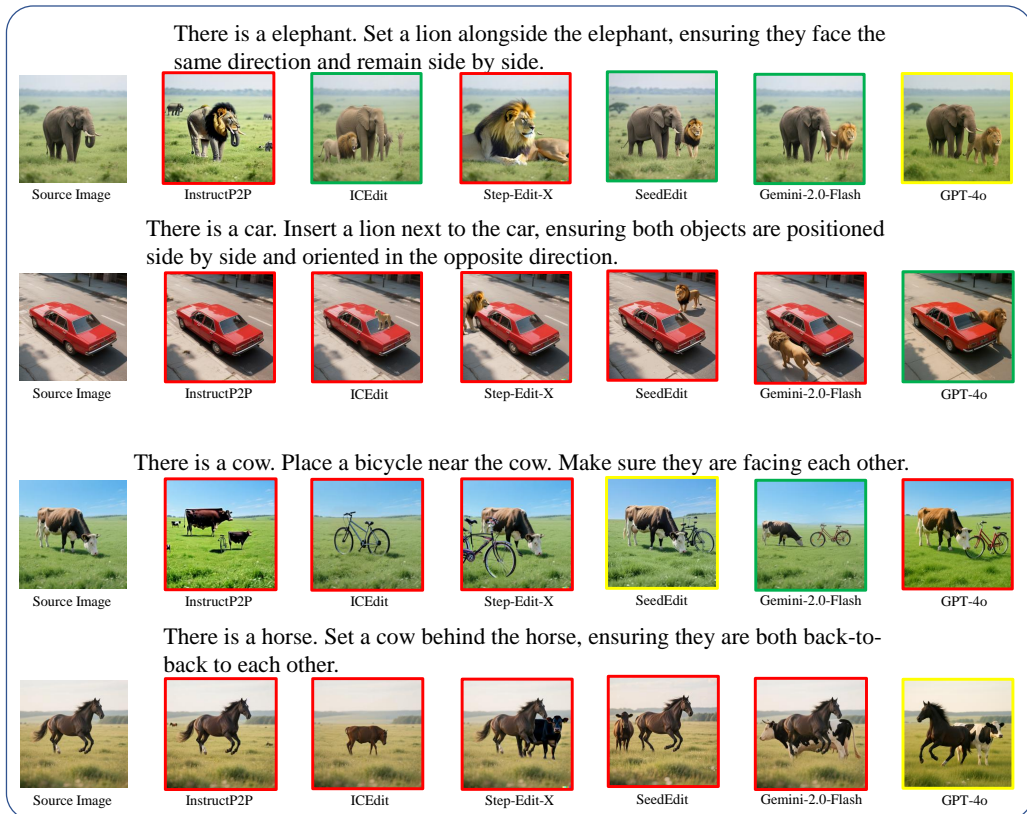


Figure 22: Visualization of Instruction-based Image Editing Benchmark on the subdomain **Intrinsic Relation**



Figure 23: Visualization of Instruction-based Image Editing Benchmark on the subdomain **Object Size**



Figure 24: Visualization of Instruction-based Image Editing Benchmark on the subdomain **Object Distance**

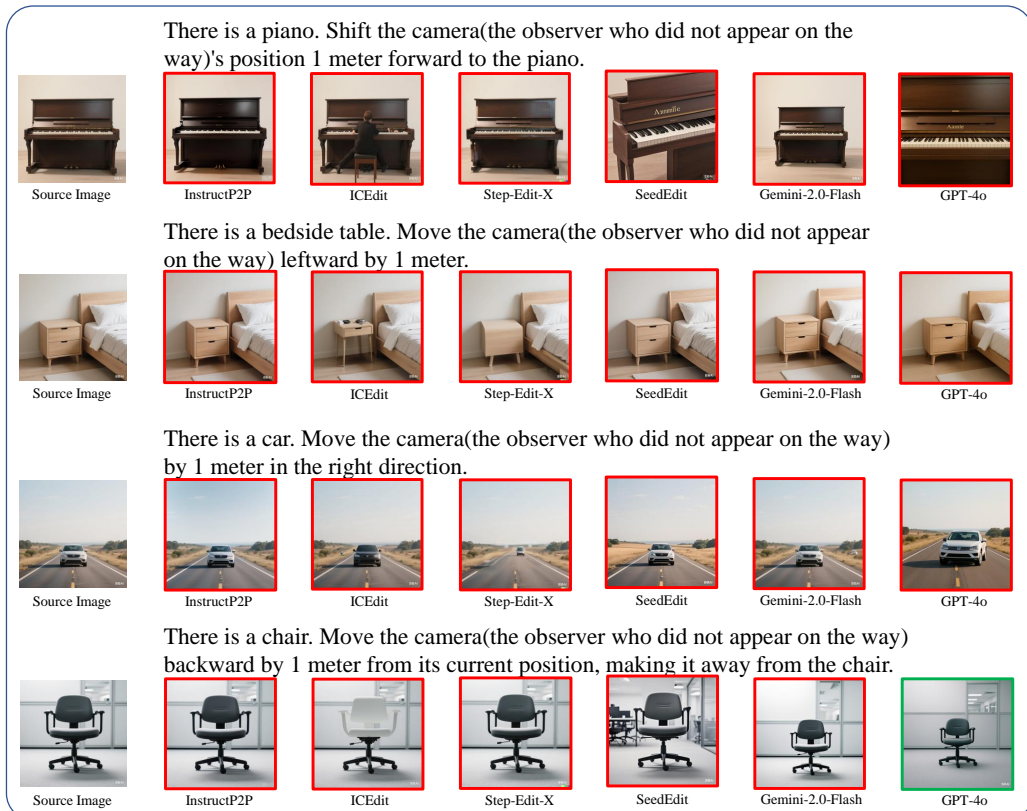


Figure 25: Visualization of Instruction-based Image Editing Benchmark on the subdomain **Camera Distance**

49 **E More Result on human alignment of different evaluators**

Evaluator	Spatial Pose			Spatial Relation			Spatial Measurement			Ave.
	Camera	Object	Complex	Ego.	Allo.	Intri.	Size	ObjDist	CamDist	
qwen-vl-max	36.0	58.0	51.0	79.0	48.0	33.0	47.0	56.0	60.0	52.00
claude-3-7-sonnet-thinking	51.0	59.0	48.0	89.0	47.0	39.0	56.0	62.0	60.0	56.78

Table 2: Human alignment of different evaluators on spatial understanding, showing their accuracy on manually labeled data.