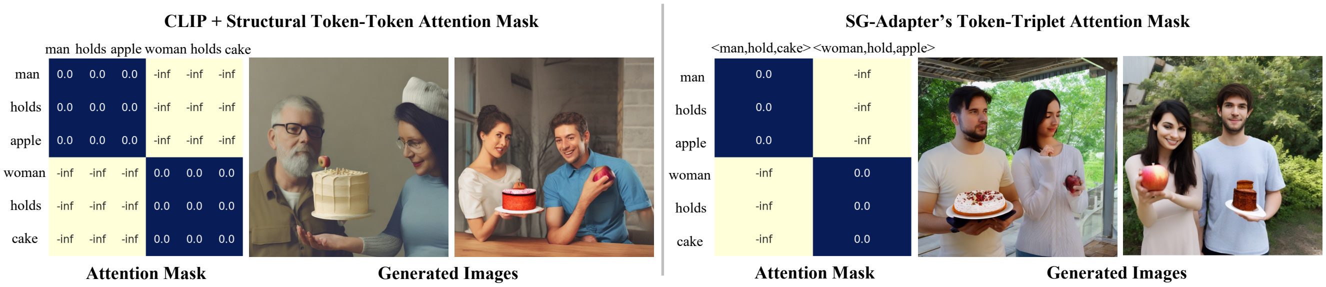
Scene Graph Guided Generation: Enable Accurate Relations Generation in Text-to-Image Models via Textural Rectification

Supplementary Material


CLIP + Structural Token-Token Attention Mask
Attention Mask
Generated Images
SG-Adapter's Token-Triplet Attention Mask
Attention Mask
Generated Images

	man	holds	apple	woman	holds	cake
man	0.0	0.0	0.0	-inf	-inf	-inf
holds	0.0	0.0	0.0	-inf	-inf	-inf
apple	0.0	0.0	0.0	-inf	-inf	-inf
woman	-inf	-inf	-inf	0.0	0.0	0.0
holds	-inf	-inf	-inf	0.0	0.0	0.0
cake	-inf	-inf	-inf	0.0	0.0	0.0

	<man,hold,cake>	<woman,hold,apple>
man	0.0	-inf
holds	0.0	-inf
apple	0.0	-inf
woman	-inf	0.0
holds	-inf	0.0
cake	-inf	0.0

Figure 6. **Results of Initial Attempt and SG-Adapter.** Directly integrating the structural token-token attention mask in a hard way within the CLIP model could break the delicate balance of learned sequential dependencies and fail to ensure accurate semantic structure in the generated images. Instead, our SG-Adapter makes use of a novel token-triplet attention mask in a learnable way to correct the unreasonable token-token interactions and guarantee precise relation correspondence.

7. Appendix

7.1. Initial Attempts with Scene Graph Attention Mask

Our initial attempts involved the utilization of an adjusted attention mask $\mathbf{M}^{\tau}$ within the transformer's attention mechanism of the CLIP text encoder, aiming to ensure intra-triplet attention cohesion and maintain the semantic structure of the caption. This integration, however, did not translate to empirical improvements as shown in Fig. 6. We observed that directly manipulating the attention mask of the pre-trained CLIP model could disrupt the delicate balance of learned sequential dependencies, adversely affecting the quality of text-to-image synthesis due to inconsistencies between the model's training and inference methodologies.

7.2. Ablation of Number of Parameters

To figure out how many parameters we need to set to SG-Adapter, We conduct an analysis of how performance varies with the size of the parameters in the adaptor. Detailed results are provided in Tab. 4. The findings show that as the number of parameters increases, FID worsens (higher values) and Alignment with Scene Graph slightly improves. We chose a balance point between FID and scene graph alignment to ensure optimal performance without over-parameterization.

number of parameters	SG-IoU	Entity-IoU	Relation-IoU	FID
8.666M	0.554	0.805	0.748	26.0
11.55M our setting	0.623	0.812	0.753	26.2
17.85M	0.628	0.803	0.766	28.2
24.15M	0.630	0.819	0.755	30.5

Table 4. Performance with different size of parameters.

7.3. Single Relation Learning

Apart from enhancing the model's ability to generate multiple relations with accurate correspondence, our method could also learn a new and complex single relation as a by-product, similar to ReVersion. Fig. 7 demonstrates SG-Adapter's ability to generate difficult single relation.

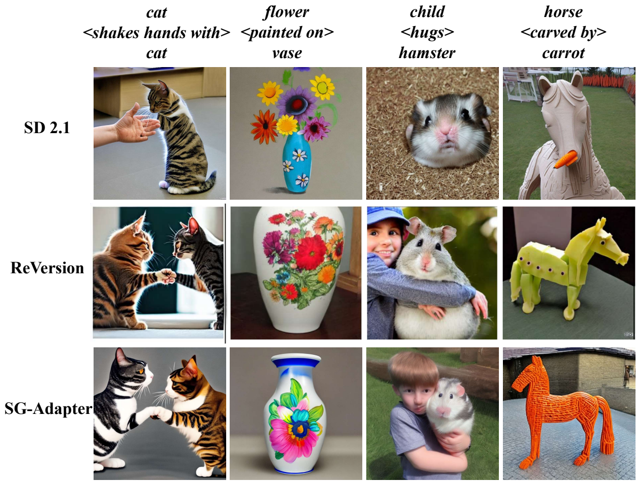


Figure 7. **The ability of SG-Adapter to learn single relation.**

7.4. MultiRels Benchmark Details

As mentioned in the main paper, we organized the MultiRels into two parts: Reversion and Multiple Relations.

This section will introduce more detailed information about the Multiple Relations part.

The Multiple Relations part contains 210 samples. We initially plan to collect all the images from the Internet and then label them manually. However, after we had collected dozens of images, we found this way not efficient enough since there are very few images that contain multiple clear and salient relations. Besides, a text describing a multiple relations image tends to be long so the relevance of retrieved images will also decrease with the longer text. Therefore we only collect 40 multi-relations images from the Internet by retrieving a long text.

On the contrary, 170 images in the Multiple Relations part were taken by ourselves.

1. Concerning human relations, we mainly focus on "*holding*", "*drinking*", "*stands on*", "*sits on*" such kinds of simple human actions, combining random 2-3 relations from them to result in 15 different text templates containing multiple relations finally. To make the collected images as diverse as possible, we arranged for 5 volunteers to participate in the photoshoot in 6 different environmental backgrounds both indoor and outdoor. **We re-paint faces in these photos locally for the privacy of the volunteers and the blind paper review**. We examples of the above data in Fig.8, Fig.9.
2. For object relations, we consider the positional relationships like "*is above*", and "*is under*". We place various normal objects, e.g. fruits and daily necessities, on the same/different side of an object like a table, chair, bench, and so on. Also, for the diversity of the data set, we take photos of these objects in 6 different indoor and outdoor environments. There are examples provided in Fig.10, Fig.11.

The 15 templates we adopt in human relations are as follows:

1. *a man stands on floor and a woman sits on a chair.*
2. *a woman stands on floor and a man sits on a chair.*
3. *a man stands on floor and a woman stands on floor.*
4. *a man sits on a chair and a woman sits on a chair.*
5. *a man drinking milk and a woman drinking cola.*
6. *a man sits on a chair and a woman holding a* $\langle obj \rangle$ *stands on floor.*
7. *a woman sits on a chair and a man holding a* $\langle obj \rangle$ *stands on floor.*
8. *a man drinking juice sits on a chair and a woman stands on floor.*
9. *a man holding a* $\langle obj \rangle$ *stands on floor and a woman drinking water stands on floor.*
10. *a man stands on floor and a man sits on a chair.*
11. *a man drinking water sits on a chair and a man holding a* $\langle obj \rangle$ *stands on floor.*
12. *a man holding a* $\langle obj \rangle$ *stands on floor and a man sits on a chair.*
13. *a man holding a* $\langle obj \rangle$ *stands on floor and a man stands on floor.*
14. *a man drinking juice and a man drinking water.*
15. *a man drinking water and a man drinking water.*

To see the complete data set and its corresponding metadata, please refer to the compressed file MultiRels.zip.

7.5. Additional Qualitative Results

We present additional qualitative results of our SG-Adapter compared with other baseline methods in Fig.12, Fig.14.

7.6. GPT-4V Prompts

Prompt to Extract Scene Graph from Image: *Please extract the scene graph of the given image. The scene graph just needs to include the relations of the salient objects and exclude the background. The scene graph should be a list of triplets like ["subject", "predicate", "object"].*
Both subject and object should be selected from the following list: ["an astronaut", "a ball", "a cake", "a box", "a television", "a table", "a horse", "a pitaya", "a woman", "a book", "a laptop", "a bottle", "a banana", "a sofa", "floor", "water", "an apple", "a chair", "a pineapple", "an umbrella", "a boy", "a paper", "a bear", "a girl", "a panda", "a cup", "a man", "a bike", "a carrot", "a phone"].
The predicate should be selected from the following list: ["stands on", "is above", "drinking", "is under", "sits back to back with", "ride on", "holding", "sits on"]
Besides the scene graph, please also output the objects list in the image like ["object1", "object2", ..., "object"]. The object should be also selected from the above-mentioned object list. The output should only contain the scene graph and the object list.

Prompt to Parse Caption to Scene Graph:*Here I have a group of captions and please help me to parse each one. Each caption should be transformed to a Scene Graph reasonably which is composed of some relations. A relation is a triplet like [subject, predicate, object] and please replace the original pronouns with reasonable nouns. Both subject and object should only have one object or person. "and" relation should not be included in the Scene Graph. Besides, I want to get the indexes of all of the subjects, predicates, and objects in the original caption, which is called mapping here.*
For example, the caption is: a boy holding a bottle shakes hands with a girl sitting on a bench.
The corresponding Scene Graph should be: [[a boy, holding, a bottle], [a boy, shakes hands with, a girl], [a girl, sitting on, a bench]].
The indexes of each word(punctuation is also considered a word) in the caption(called all_words_indexes) are:"a":0, "boy":1, "holding":2, "a" :3, "bottle":4, "shakes":5, "hands":6, "with":7, "a":8, "girl":9, "sit-

ting":10, "on":11, "a":12, "bench":13.
The index of every word in the Scene Graph(i.e., the mapping) should be:["a":0, "boy":1, "holding":2," a":3, "bottle":4,"a":0, "boy":1, "shakes":5, "hands":6, "with":7, "a":8, "girl":9, "sitting":10, "on": 11,"a":12, "bench":13]
The indexes of all of the subject, predicate, and object in the original caption should be highly precise. The results of each caption are data like:
scene graph:
all_words_indexes:
mapping:
When generating the mapping please refer to the scene graph and the all_words_indexes to ensure the correct result.
The captions are:

7.7. Implementation Details

All the experiments are conducted on 768×768 image resolution and all the models are trained on a single A100 GPU for several hours. We adopt SD 2.1 as our base model. We set training batch size 4, learning rate 1e-5. We adopt the optimizer AdamW and most models converge around 12000 to 14000 iterations. During the sampling process, we use a parameter τ to balance the ability to control and the image quality. For a diffusion process with T steps, we could use scene graph guided inference at the first $\tau * T$ steps and use the standard inference for the remaining $(1 - \tau) * T$ steps. In this paper, τ is set to 0.3, which is enough to achieve relation control while maintaining the image quality.

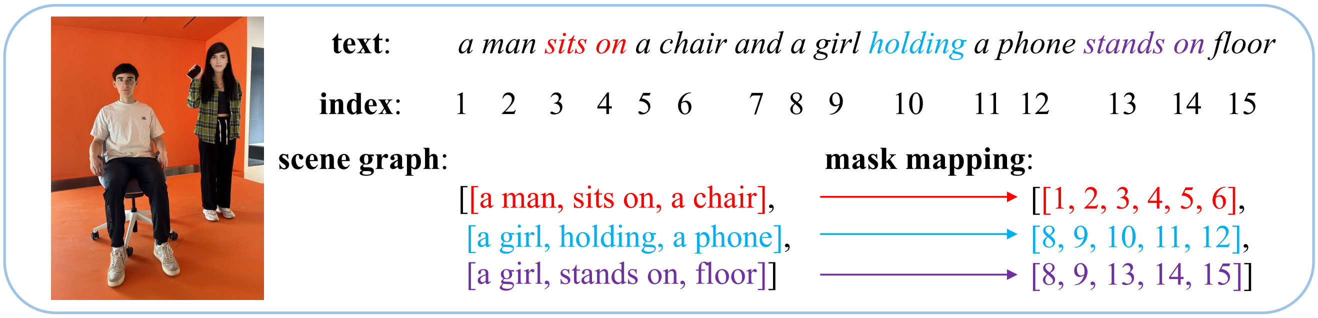

text: a man sits on a chair and a girl holding a phone stands on floor
index: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
scene graph:
[[a man, sits on, a chair],
[a girl, holding, a phone],
[a girl, stands on, floor]]
mask mapping:
[[1, 2, 3, 4, 5, 6],
[8, 9, 10, 11, 12],
[8, 9, 13, 14, 15]]

Figure 8. **Example 1 of MultiRels**

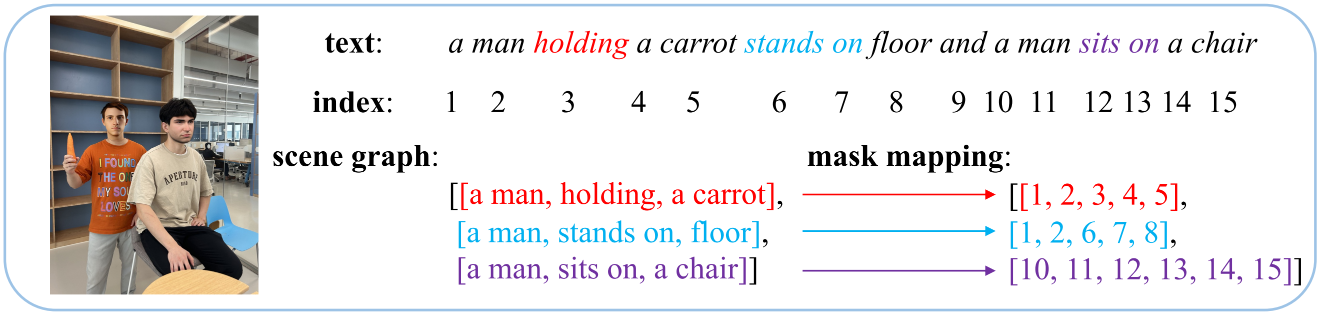

text: a man holding a carrot stands on floor and a man sits on a chair
index: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
scene graph:
[[a man, holding, a carrot],
[a man, stands on, floor],
[a man, sits on, a chair]]
mask mapping:
[[1, 2, 3, 4, 5],
[1, 2, 6, 7, 8],
[10, 11, 12, 13, 14, 15]]

Figure 9. **Example 2 of MultiRels**

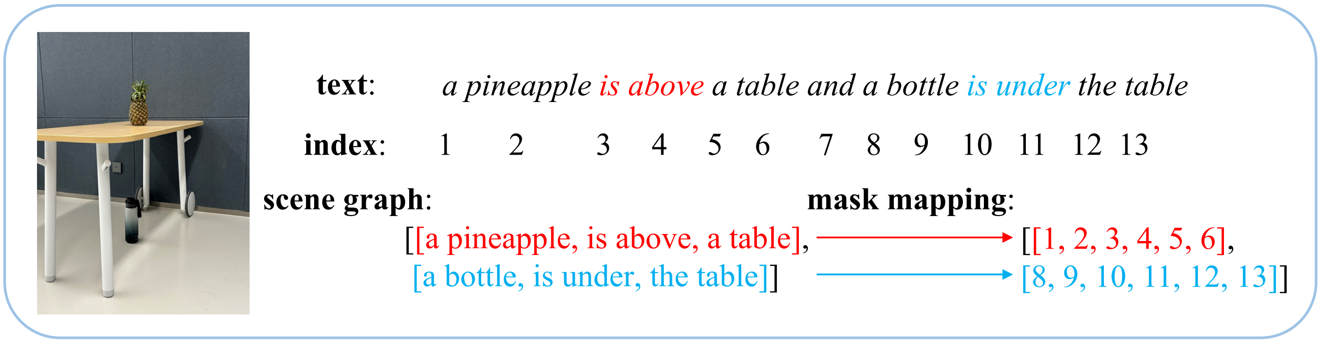

text: a pineapple is above a table and a bottle is under the table
index: 1 2 3 4 5 6 7 8 9 10 11 12 13
scene graph:
[[a pineapple, is above, a table],
[a bottle, is under, the table]]
mask mapping:
[[1, 2, 3, 4, 5, 6],
[8, 9, 10, 11, 12, 13]]

Figure 10. **Example 3 of MultiRels**

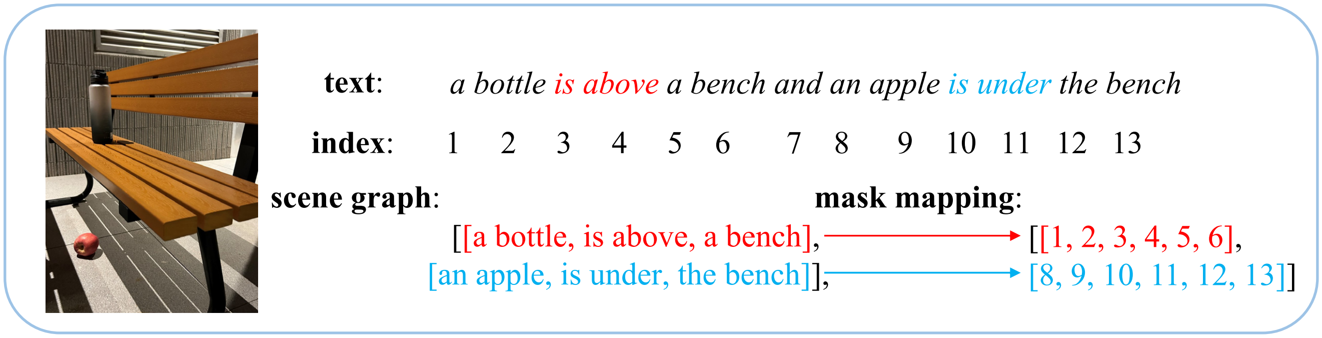

text: a bottle is above a bench and an apple is under the bench
index: 1 2 3 4 5 6 7 8 9 10 11 12 13
scene graph:
[[a bottle, is above, a bench],
[an apple, is under, the bench]],
mask mapping:
[[1, 2, 3, 4, 5, 6],
[8, 9, 10, 11, 12, 13]]

Figure 11. **Example 4 of MultiRels**

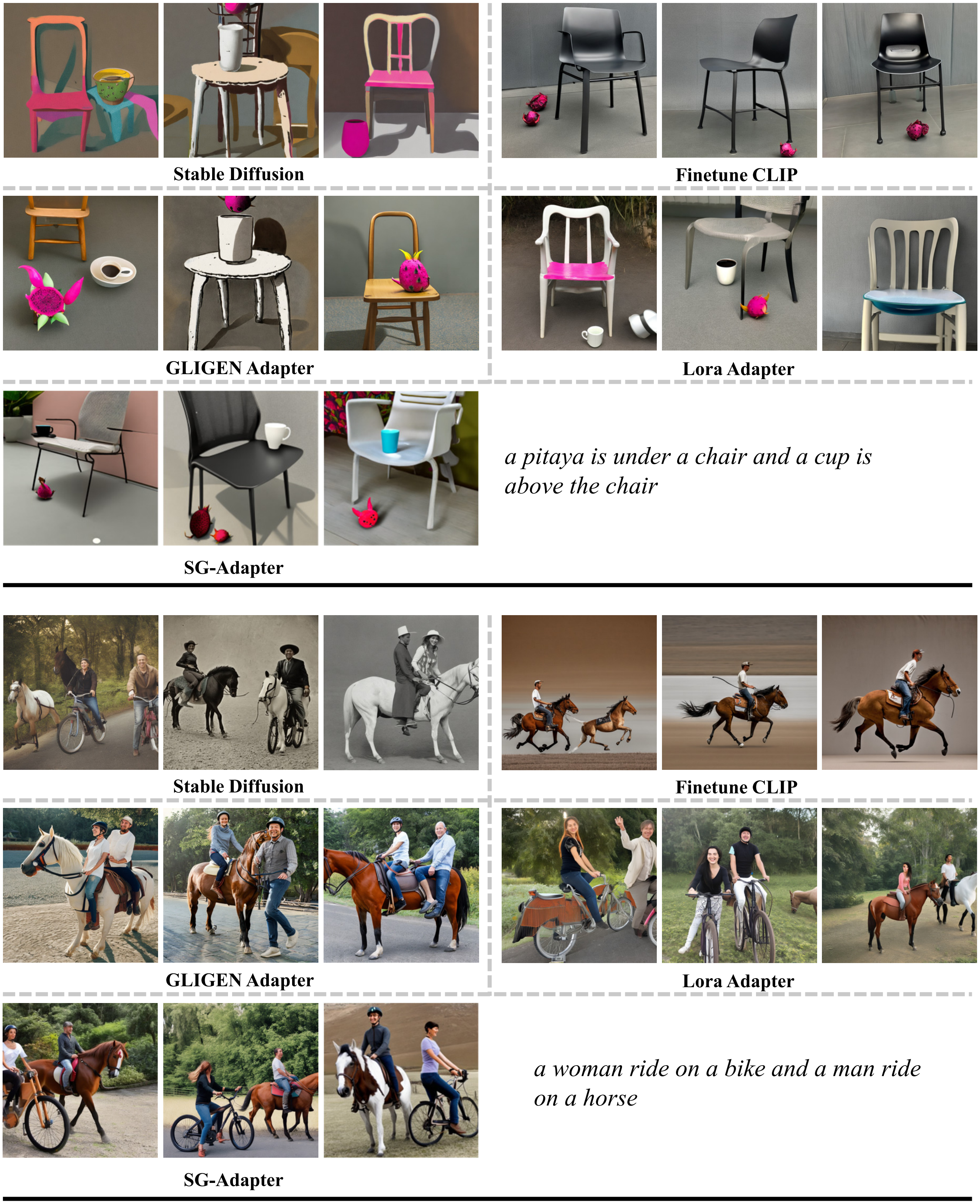


Figure 12. **More qualitative results-1.**

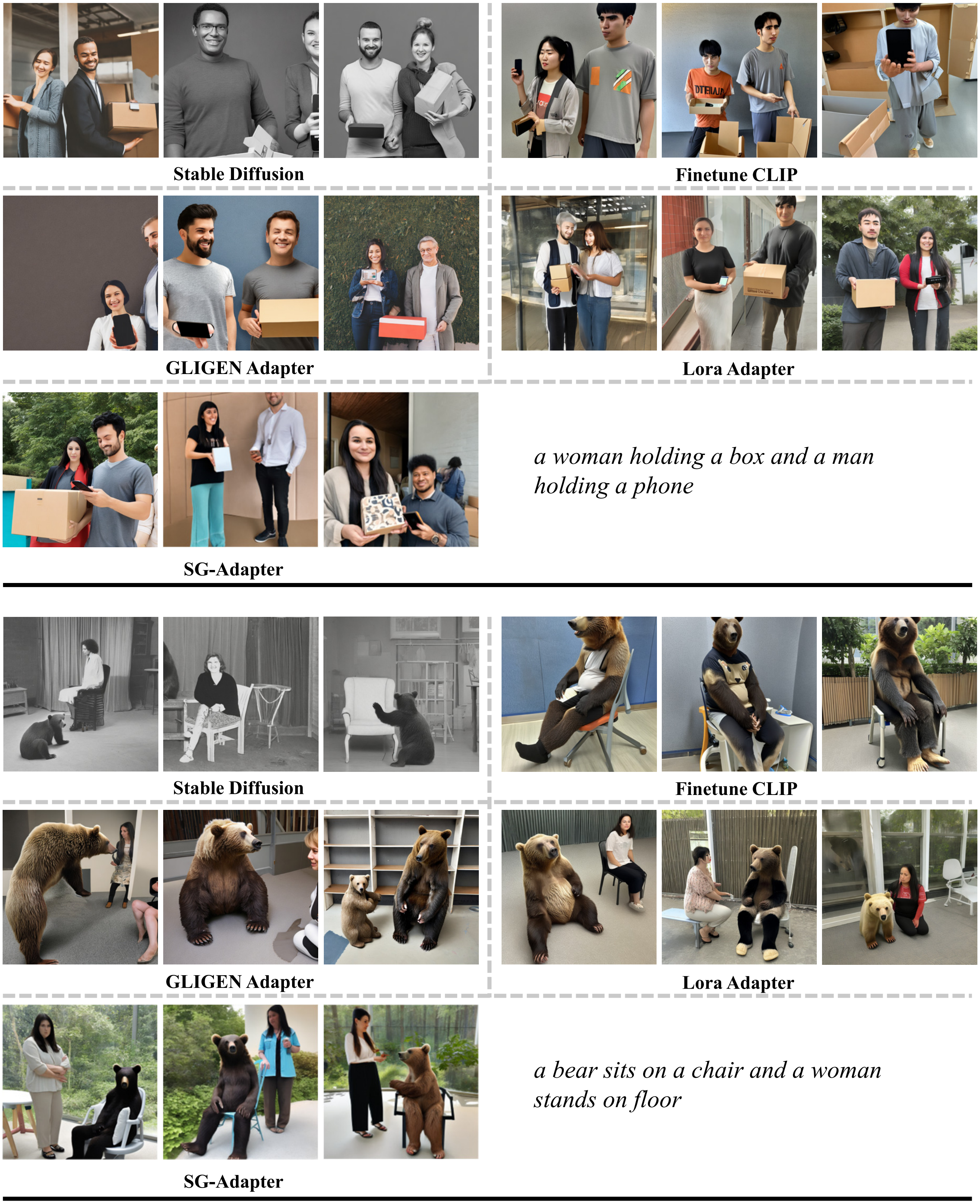


Figure 13. **More qualitative results-2.**

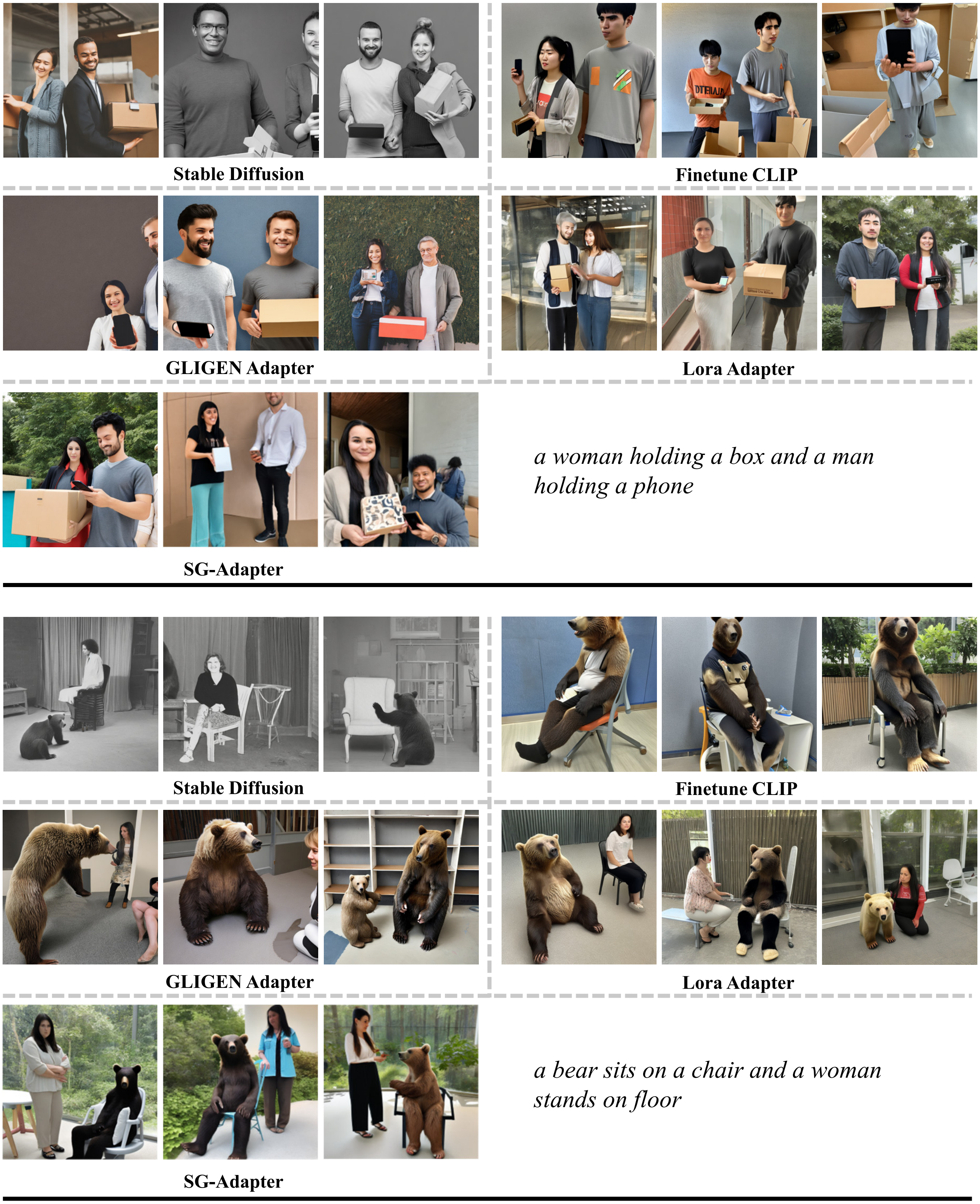


Figure 14. **More qualitative results-2.**