Check, Locate, Rectify: A Training-Free Layout Calibration System for Text-to-Image Generation

Supplementary Material

A. Relation Vocabulary for Checking

Our SimM determines the existence of layout requirements by checking whether any words from our predefined relation vocabulary are present in the prompt. According to the semantic similarity, the vocabulary contains six categories:

- **left**: "*left*", "*west*"
- **right**: "*right*", "*east*"
- **above**: "*above*", "*over*", "*on*", "*top*", "*north*"
- **below**: "*below*", "*beneath*", "*underneath*", "*under*", "*bottom*", "*south*"
- **between**: "*between*", "*among*", "*middle*"
- **additional superlative**: "*upper-left*", "*upper-right*", "*lower-left*", "*lower-right*"

Note that (1) The "additional superlative" category serves as a supplement for words that have not been covered. In the given context, words such as "*left*" and "*above*" can also represent the superlative relations. (2) This vocabulary can easily be extended according to the needs of the dataset.

B. Superlative Predefined Positions

For each object associated with a superlative relation, the relative bounding box $\widehat{\mathbf{b}} = (\widehat{x}, \widehat{y}, \widehat{w}, \widehat{h})$ is assigned as follows:

- **left**: (0.20, 0.50, 0.33, 1.00)
- **right**: (0.80, 0.50, 0.33, 1.00)
- **above**: (0.50, 0.20, 1.00, 0.33)
- **below**: (0.50, 0.80, 1.00, 0.33)
- **middle**: (0.50, 0.50, 0.50, 0.50)
- **upper-left**: (0.25, 0.25, 0.50, 0.50)
- **upper-right**: (0.75, 0.25, 0.50, 0.50)
- **lower-left**: (0.25, 0.75, 0.50, 0.50)
- **lower-right**: (0.75, 0.75, 0.50, 0.50)

Table 2. **Detailed quantitative results on SimMBench.** The generation accuracy (%) is reported.

Methods	1 object	2 objects	3 objects	4 objects
Stable Diffusion [32]	15.56	5.21	0.00	0.00
BoxDiff [40]	41.11	18.23	19.64	13.33
Layout-Guidance [6]	82.22	5.73	3.57	20.00
Attention-Refocusing [25]	65.56	41.67	57.14	53.33
SimM (Ours)	**82.22**	**53.64**	**76.79**	**66.67**

C. An Example of Target Layout Generation

To facilitate understanding of how SimM parses the prompt and generates the target bounding box for each object with a set of heuristic rules, we show an example in Fig. 9 to illustrate it more clearly. Specifically, the process can be roughly divided into four steps:

1. Semantic parsing. SimM parses the superlative tuples and relative triplets from the prompt. And the relative triples can be organized as a semantic tree, with nodes as objects and edges as spatial relations.

2. Assign the superlative boxes. Given each superlative tuple, SimM assigns a predefined target box to the object according to its superlative position term.

3. Traverse the semantic tree for a global view. By traversing the tree, SimM organizes the global layout of the remaining objects.

4. Assign the relative boxes. SimM allocates the remaining space to the objects associated with superlative relations.

D. Benchmark Details

Overview. Our proposed SimMBench focuses on superlative relations. Specifically, to sample an evaluation prompt, we first determine the number of objects in the prompt. Each prompt contains a minimum of one object and a maximum of four objects. Then, we sample the superlative relation for each object that has not yet been determined, where

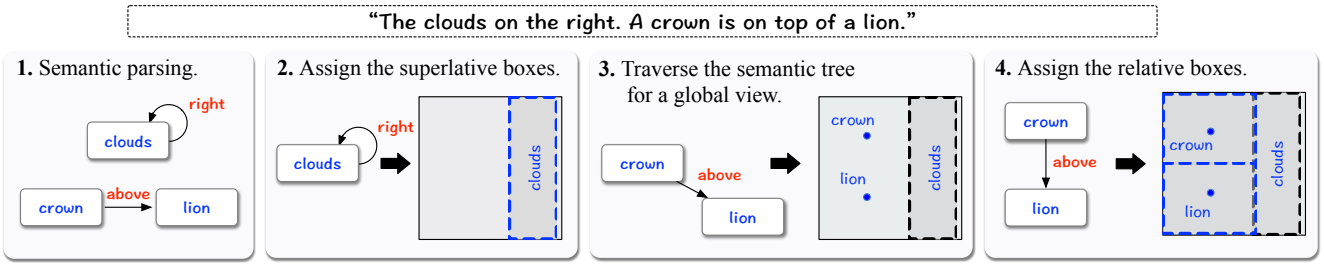

Figure 9. **Example of target layout generation.**

Figure 10. **Qualitative results on HRS [2] and TIFA [18] benchmarks.**

the predefined superlative relation set is the same as shown in Appendix B. Finally, we sample the objects present in the current prompt from a predefined set of objects. To better evaluate the impact of layout requirements on image generation, a sampled object set can be shared between prompts with different superlative relations. As a result, SimMBench contains 203 different prompts. The number of prompts containing 1/2/3/4 objects is 36/96/56/15. And the number of occurrences of each superlative relation is 55/55/49/49/56/48/48/48/48. The benchmark will be publicly available.

Object set. The predefined object set consists of 28 different items as follows:

- **single-word**: "*backpack*", "*flower*", "*crown*", "*towel*", "*scarf*", "*beach*", "*clouds*", "*tree*", "*table*", "*book*", "*handbag*", "*bus*", "*bicycle*", "*car*", "*motorcycle*", "*cat*", "*dog*", "*horse*"
- **phrase**: "*chocolate cookie*", "*strawberry cake*", "*vanilla ice cream cone*"
- **with color**: "*yellow sunflower*", "*gray mountain*", "*white daisy*", "*pink cupcake*", "*red tomato*", "*golden saxophone*", "*green broccoli*"

E. Detailed Accuracies on SimMBench

In Tab. 2, we report the accuracies when the number of objects in the prompt is different. It can be observed that our SimM outperforms the baselines in all cases. Furthermore, despite the simplicity of the case with a single object, the accuracies do not show a clear downward trend as the number of objects increases. The difficulty of accurately representing the layout is also influenced by the specific layout requirements of the objects and their context.

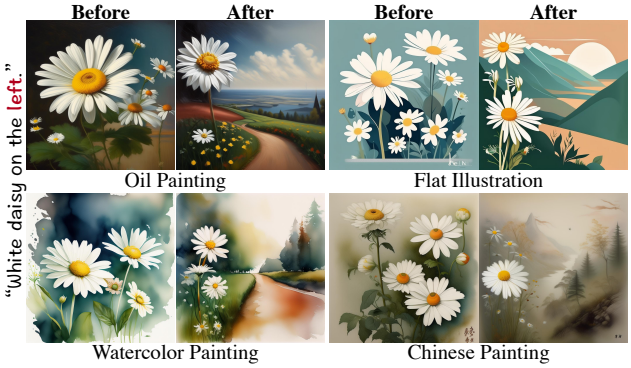

Figure 11. **Layout calibration results of images in different styles.**

F. Additional Results

F.1. Latency Comparison for Layout Generation

Since our SimM system presents a new solution for generating the target layout, we provide a brief discussion of the observed increase in latency here. Existing layout-to-image works [25, 27] commonly rely on GPT-4 [24], however, each invocation of the API requires a response time of ∼3 seconds. In contrast, thanks to the industrial-strength library, our proposed solution requires an average of only 0.006 seconds for each prompt and does not require a GPU. This significantly improves the user experience for real-time text-to-image generators.

F.2. Generalization Across Diverse Styles

In practical scenarios, users often request the text-to-image generators to produce images in specific styles. In Fig. 11, we show that the stylistic demands for generated images do not hinder the rectification of the layout by SimM.

F.3. Qualitative Results on Other Benchmarks

We additionally present the qualitative results obtained on two latest benchmarks, HRS [2] and TIFA [18], in Fig. 10. These two benchmarks, similar to DrawBench, excessively focus on relative spatial relations. Due to the cost of comprehensive manual evaluation, we take quantitative evaluation on these benchmarks as future work.

F.4. Comparison with Training-Based Method

LayoutDiffusion [47] is a representative approach in training auxiliary modules to embed the layout information into intermediate features for controlling. However, it is constrained to fixed categories, thereby rendering it unsuitable for various datasets including Drawbench. To compare our SimM with LayoutDiffusion, we select prompts that only includes valid objects for LayoutDiffusion from our SimM-Bench. As observed in Fig. 12, the limitation of layout significantly reduces the generation quality of LayoutDiffusion, resulting in its performance being far inferior to SimM.

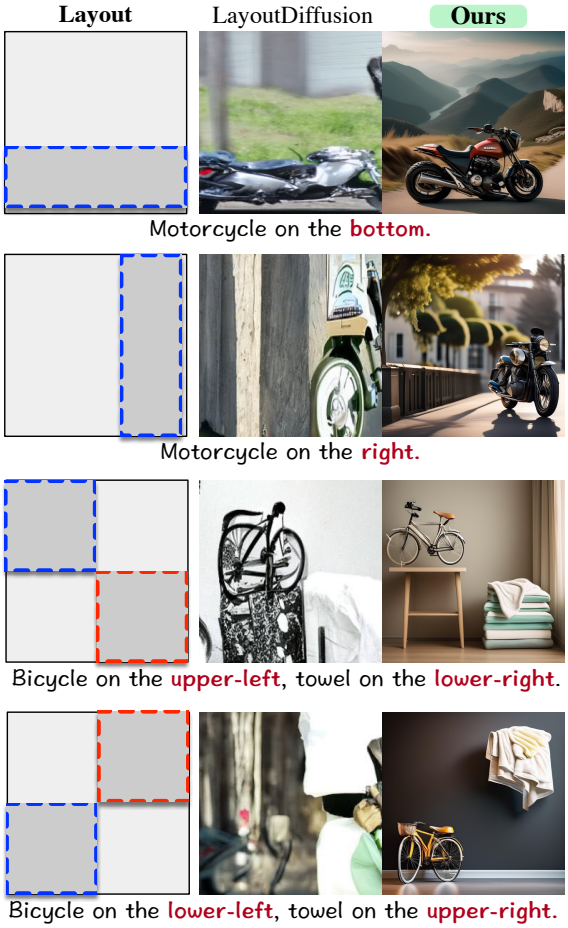

Figure 12. **Qualitative comparisons with LayoutDiffusion [47].** The generation quality of LayoutDiffusion is far worse than SimM.

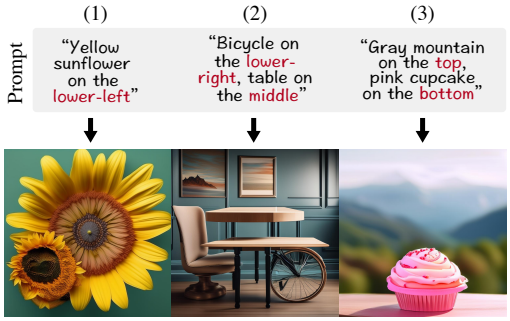

Figure 13. **Typical failure cases identified by human evaluators.**

F.5. Failure Case Analysis

In Fig. 13, we present typical cases of what the human evaluators perceive as errors. The first case is the repeated generation of objects with some in the wrong position. The second case is that multiple objects interact with each other during generation, resulting in incomplete generation. The third case includes missing or unclear objects. These errors are mostly due to the fact that a single adjustment strength parameter α may not be optimal for all generation. This results in insufficient activation enhancement or suppression on the attention map, leading to inaccuracies in generating all objects accurately or preventing the repeated generation.

F.6. Additional Qualitative Comparison Results

To show the effectiveness of SimM, we illustrate additional qualitative results in Figs. 14 and 15.

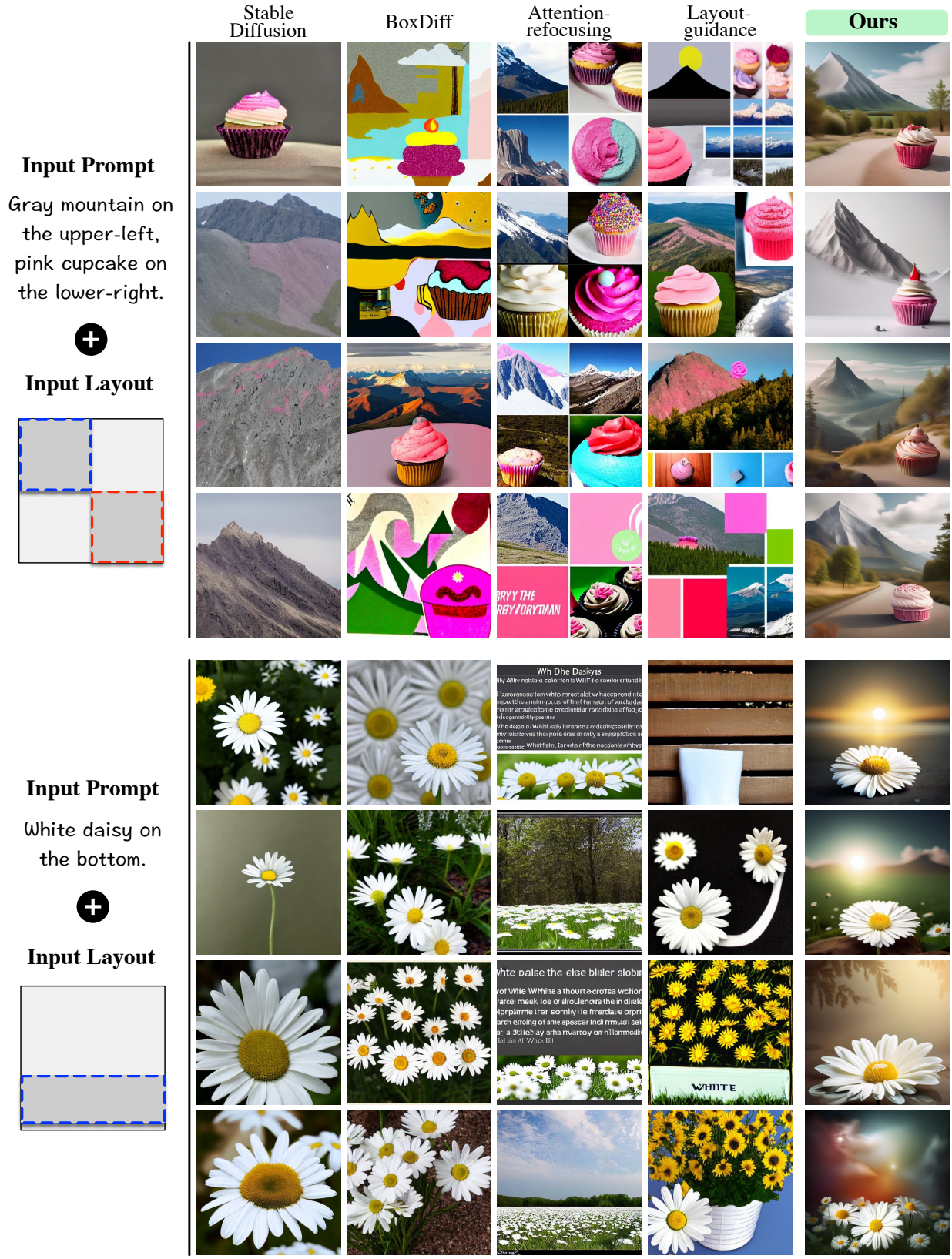

Figure 14. **Additional qualitative comparisons.**

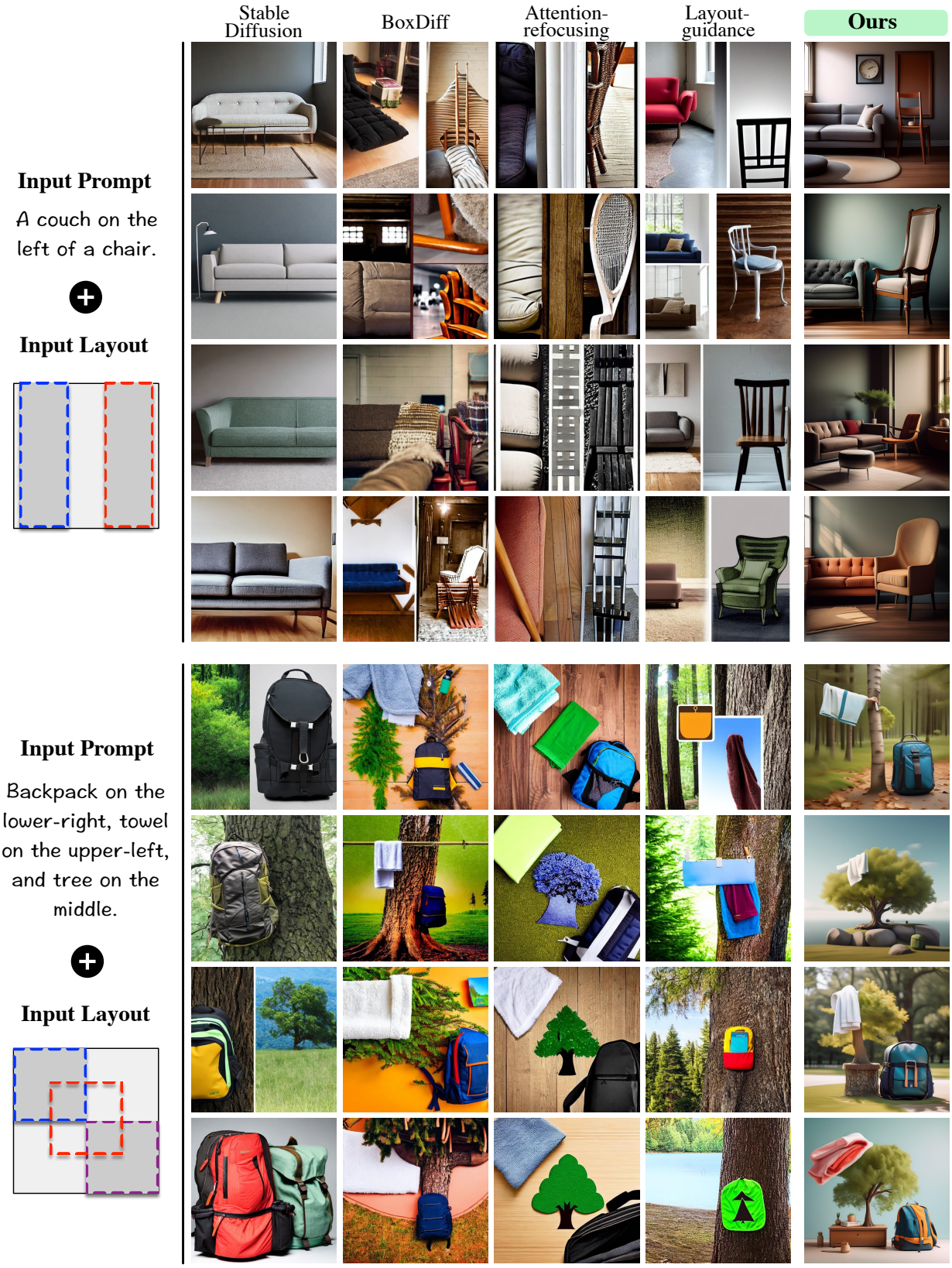

Figure 15. **Additional qualitative comparisons.**