A Implementation Details

Evaluation details. We employ pre-trained models for our proposed evaluation metrics. For BLIP-VQA, we utilize the BLIP w/ ViT-B and CapFilt-L [1] pretrained on image-text pairs and fine-tuned on VQA. We employ the UniDet [2] model Unified_learned_COIM_RS200_6x+2x trained on 4 large-scale detection datasets (COCO [3], Objects365 [4], OpenImages [5], and Mapillary [6]). For CLIPScore, we use the "ViT-B/32" pretrained CLIP model [7, 8]. Finally, for MiniGPT4-CoT, we utilize the Vicuna 13B of MiniGPT4 [9] variant with a temperature setting of 0.7 and a beam size of 1.

Training details. We implement our proposed FT-SSWL upon the codebase of diffusers [10] (Apache License), and finetune the self-attention layers of the CLIP text encoder and the attention layers of U-net using LoRA [11]. The model is trained by AdamW optimizer [12] with β_1=0.9, β_2=0.999, ϵ=1e-8, and weight decay of 0.01. The batch size is 5. The model is trained on 8 32GB NVIDIA v100 GPUs, for 50000-100000 steps.

B T2I-CompBench Dataset Construction

This section provides the details of the prompts that ChatGPT uses for generating the text prompts in T2I-CompBench. The text prompts in T2I-CompBench is available at this link.

Color. The prompt for ChatGPT is: *Please generate prompts in the format of "a {adj} {noun} and a {adj} {noun}" by using the color adj. , such as "a green bench and a red car".*

Shape. (1) For fixed sentence template, the prompt for ChatGPT is: *Please generate prompts in the format of "a {adj} {noun} and a {adj} {noun}" by using the shape adj.: long, tall, short, big, small, cubic, cylindrical, pyramidal, round, circular, oval, oblong, spherical, triangular, square, rectangular, conical, pentagonal, teardrop, crescent, and diamond.* (2) For natural prompts, the prompt for ChatGPT is: *Please generate objects with shape adj. in a natural format by using the shape adj.: long, tall, short, big, small, cubic, cylindrical, pyramidal, round, circular, oval, oblong, spherical, triangular, square, rectangular, conical, pentagonal, teardrop, crescent, and diamond.*

Texture. (1) We generate 200 natural text prompts by ChatGPT with the following prompt: *Please generate objects with texture adj. in a natural format by using the texture adj.: rubber, plastic, metallic, wooden, fabric, fluffy, leather, glass.* (2) Besides the ChatGPT-generated text prompts, we also provide the predefined texture attributes and objects that can be described by each texture, as shown in Table 1. We generate 800 text prompts by randomly selecting from the possible combinations of two objects each associated with a textural attribute, *e.g.*, "A rubber ball and a plastic bottle".

Table 1: Textural attributes and associated objects to construct the attribute-texture prompts.

Textures	Objects
Rubber	band, ball, tire, gloves, sole shoes, eraser, boots, mat
Plastic	Bottle, bag, toy, cutlery, chair, phone case, container, cup, plate
Metallic	car, jewelry, watch, keychain, desk lamp, door knob, spoon, fork, knife, key, ring, necklace, bracelet, earring
Wooden	chair, table, picture frame, toy, jewelry box, door, floor, chopsticks, pencils, spoon, knife
Fabric	bag, pillow, curtain, shirt, pants, dress, blanket, towel, rug, hat, scarf, sweater, jacket
Fluffy	pillow, blanket, teddy bear, rug, sweater, clouds, towel, scarf, hat
Leather	jacket, shoes, belt, bag, wallet, gloves, chair, sofa, hat, watch
Glass	bottle, vase, window, cup, mirror, jar, table, bowl, plate

Non-spatial relation. The prompt for ChatGPT is: *Please generate natural prompts that contain subjects and objects by using relationship words such as wear, watch, speak, hold, have, run, look at, talk to, jump, play, walk with, stand on, and sit on.*

Complex. (1) For 2 objects with mixed attributes, the prompt for ChatGPT is: *Please generate natural compositional phrases, containing 2 objects with each object one adj. from {color, shape, texture} descriptions and spatial (left/right/top/bottom/next to/near/on side of) or non-spatial relationships.*

(2) For 2 objects with multiple attributes, the prompt for ChatGPT is: *Please generate natural compositional phrases, containing 2 objects with several adj. from {color, shape, texture} descriptions and spatial (left/right/top/bottom/next to/near/on side of) or non-spatial relationships.* (3) For multiple objects with mixed attributes, the prompt for ChatGPT is: *Please generate natural compositional phrases, containing multiple objects (number>2) with each one adj. from {color, shape, texture} descriptions and spatial (left/right/top/bottom/next to/near/on side of) non-spatial relationships.* (4) For multiple objects with multiple attributes, the prompt for ChatGPT is: *Please generate natural compositional phrases, containing multiple objects (number>2) with several adj. from {color, shape, texture} descriptions and spatial (left/right/top/bottom/next to/near/on side of) or non-spatial relationships.*

C Evaluation Metrics

C.1 Prompts for MiniGPT4-CoT and MiniGPT4 Evaluation

MiniGPT4-Chain-of-Thought. In this part, we detail the prompts used for the MiniGPT4-CoT evaluation metric. For each sub-category, we ask two questions in sequence: "describe the image" and "predict the image-text alignment score". Specifically, Table 2 shows the MiniGPT4-CoT prompts for evaluating attribute binding (color, shape, texture). Table 3, Table 4, and Table 5 demonstrate the prompt templates used for spatial relationships, non-spatial relationships, and complex compositions, respectively.

Table 2: Prompts details for mGPT4-CoT evaluation on attribute binding.

Describe	You are my assistant to identify any objects and their color (shape, texture) in the image. Briefly describe what it is in the image within 50 words.
Predict	According to the image and your previous answer, evaluate if there is {adj.+noun} in the image. Give a score from 0 to 100, according the criteria: 100: there is {noun}, and {noun} is {adj}. 75: there is {noun}, {noun} is mostly {adj}. 20: there is {noun}, but it is not {adj}. 10: no {noun} in the image. Provide your analysis and explanation in JSON format with the following keys: score (e.g., 85), explanation (within 20 words).

Table 3: Prompts details for mGPT4-CoT evaluation on spatial relationship.

Describe	You are my assistant to identify objects and their spatial layout in the image. Briefly describe the image within 50 words.
Predict	According to the image and your previous answer, evaluate if the text "{xxx}" is correctly portrayed in the image. Give a score from 0 to 100, according the criteria: 100: correct spatial layout in the image for all objects mentioned in the text. 80: basically, spatial layout of objects matches the text. 60: spatial layout not aligned properly with the text. 40: image not aligned properly with the text. 20: image almost irrelevant to the text. Provide your analysis and explanation in JSON format with the following keys: score (e.g., 85), explanation (within 20 words).

MiniGPT-4 without Chain-of-Thought To guide miniGPT4 in addressing specific compositional problems, we utilize predefined prompts that prompt miniGPT4 to provide a score ranging from 0 to 100. For attribute binding, we focus on the presence of specific objects and their corresponding attributes. We utilize a prompt template such as *"Is there {object} in the image? Give a score from 0 to 100. If {object} is not present or if {object} is not {color/shape/texture description}, give a lower score."* We leverage this question for each noun phrase in the text and compute the average score. For the spatial relationships, non-spatial relationships, and complex compositions, we employ a more general prompt template such as *"Rate the overall alignment between the image and the text prompt {prompt}. Give a score from 0 to 100."*.

Table 4: Prompts details for mGPT4-CoT evaluation on non-spatial relationship.

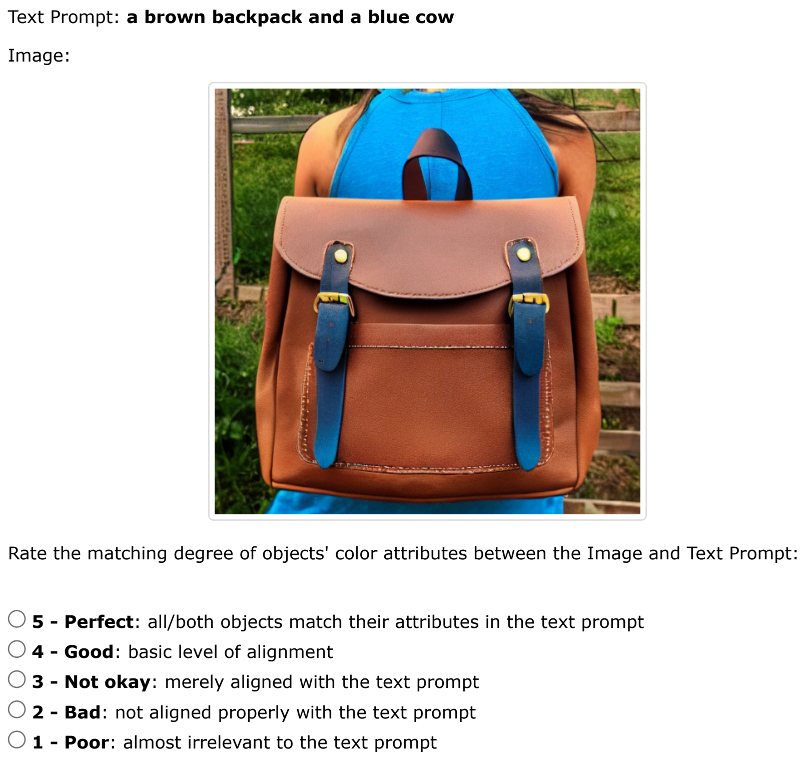
Describe	You are my assistant to identify the actions, events, objects and their relationships in the image. Briefly describe the image within 50 words.
Predict	According to the image and your previous answer, evaluate if the text "{xxx}" is correctly portrayed in the image. Give a score from 0 to 100, according the criteria: 100: the image accurately portrayed the actions, events and relationships between objects described in the text. 80: the image portrayed most of the actions, events and relationships but with minor discrepancies. 60: the image depicted some elements, but action relationships between objects are not correct. 40: the image failed to convey the full scope of the text. 20: the image did not depict any actions or events that match the text. Provide your analysis and explanation in JSON format with the following keys: score (e.g., 85), explanation (within 20 words).

Table 5: Prompts details for mGPT4-CoT evaluation on complex compositions.

Describe	You are my assistant to evaluate the correspondence of the image to a given text prompt. Briefly describe the image within 50 words, focus on the objects in the image and their attributes (such as color, shape, texture), spatial layout and action relationships.
Predict	According to the image and your previous answer, evaluate how well the image aligns with the text prompt: {xxx}. Give a score from 0 to 100, according the criteria: 100: the image perfectly matches the content of the text prompt, with no discrepancies. 80: the image portrayed most of the actions, events and relationships but with minor discrepancies. 60: the image depicted some elements in the text prompt, but ignored some key parts or details. 40: the image did not depict any actions or events that match the text. 20: the image failed to convey the full scope in the text prompt. Provide your analysis and explanation in JSON format with the following keys: score (e.g., 85), explanation (within 20 words).

C.2 Human Evaluation

We conducted human evaluations on Amazon Mechanical Turk (AMT). Specifically, we ask the annotators to rate the alignment between a generated image and the text prompt used to generate the image. Figure 1, 2, 3, 4, 5, 6 show the interfaces for human evaluation over the 6 sub-categories. We randomly sample 25 prompts from each sub-category and each model, and then randomly select 2 images per prompt. In total, we gather $1,800$ text-image pairs for human evaluation experiments. Each image-text pair is rated by 3 human annotators with a score from 1 to 5 according to the image-text alignment. The estimated hourly wage paid to each participant is 9 USD. We spend 270 USD in total on participant compensation.



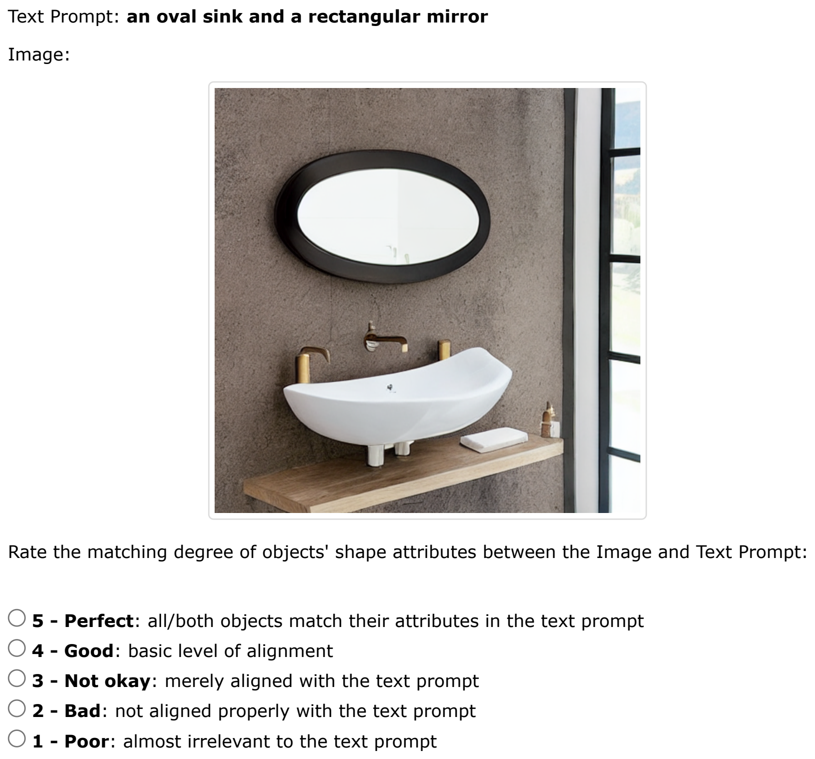
Text Prompt: **a brown backpack and a blue cow**

Image:

Rate the matching degree of objects' color attributes between the Image and Text Prompt:

○ **5 - Perfect**: all/both objects match their attributes in the text prompt
○ **4 - Good**: basic level of alignment
○ **3 - Not okay**: merely aligned with the text prompt
○ **2 - Bad**: not aligned properly with the text prompt
○ **1 - Poor**: almost irrelevant to the text prompt

Figure 1: AMT Interface for the image-text alignment evaluation on attribute binding (color).



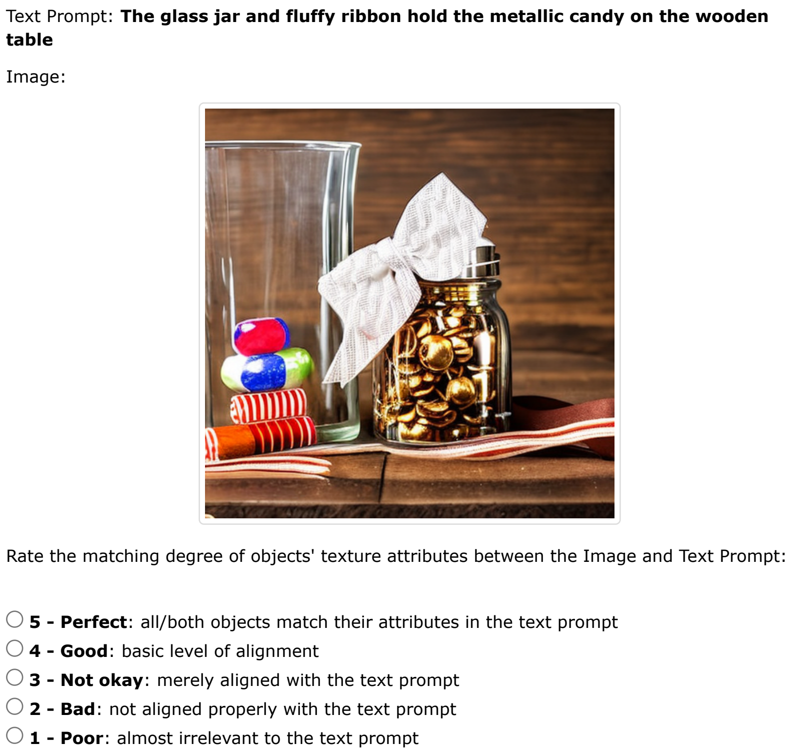
Text Prompt: **an oval sink and a rectangular mirror**

Image:

Rate the matching degree of objects' shape attributes between the Image and Text Prompt:

○ **5 - Perfect**: all/both objects match their attributes in the text prompt
○ **4 - Good**: basic level of alignment
○ **3 - Not okay**: merely aligned with the text prompt
○ **2 - Bad**: not aligned properly with the text prompt
○ **1 - Poor**: almost irrelevant to the text prompt

Figure 2: AMT Interface for the image-text alignment evaluation on attribute binding (shape).



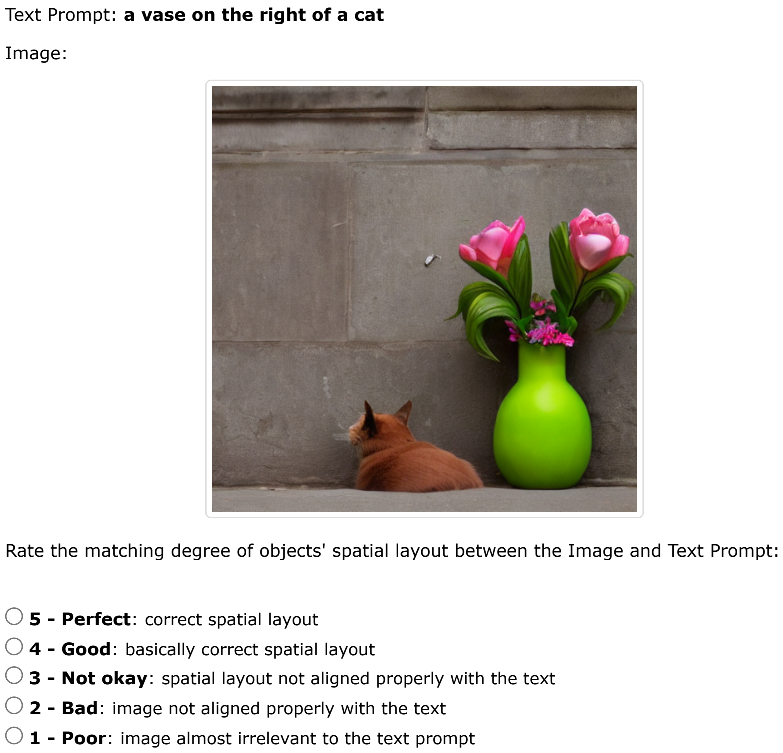
Text Prompt: **The glass jar and fluffy ribbon hold the metallic candy on the wooden table**

Image:

Rate the matching degree of objects' texture attributes between the Image and Text Prompt:

○ **5 - Perfect**: all/both objects match their attributes in the text prompt
○ **4 - Good**: basic level of alignment
○ **3 - Not okay**: merely aligned with the text prompt
○ **2 - Bad**: not aligned properly with the text prompt
○ **1 - Poor**: almost irrelevant to the text prompt

Figure 3: AMT Interface for the image-text alignment evaluation on attribute binding (texture).



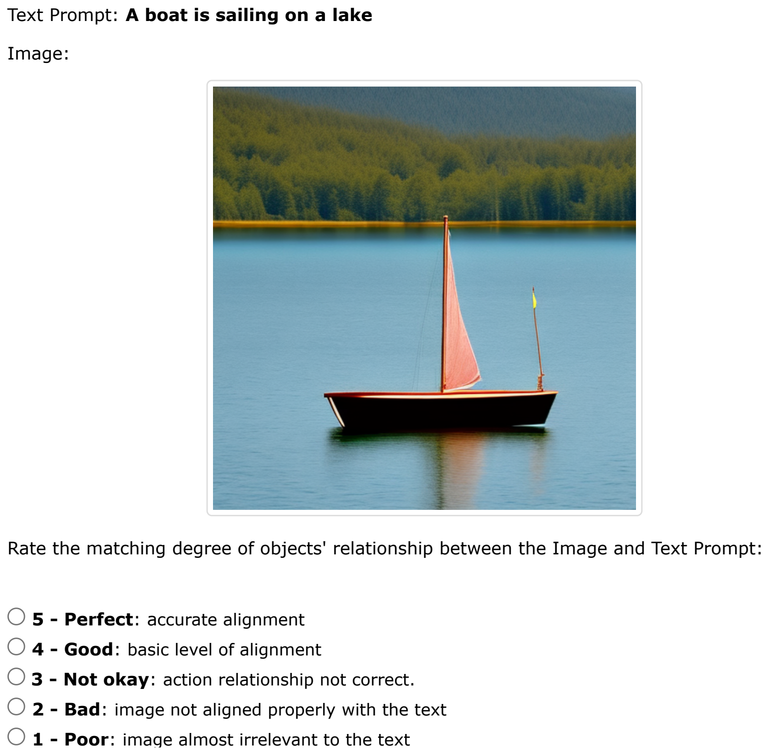
Text Prompt: **a vase on the right of a cat**

Image:

Rate the matching degree of objects' spatial layout between the Image and Text Prompt:

○ **5 - Perfect**: correct spatial layout
○ **4 - Good**: basically correct spatial layout
○ **3 - Not okay**: spatial layout not aligned properly with the text
○ **2 - Bad**: image not aligned properly with the text
○ **1 - Poor**: image almost irrelevant to the text prompt

Figure 4: AMT Interface for the image-text alignment evaluation on spatial relationships.



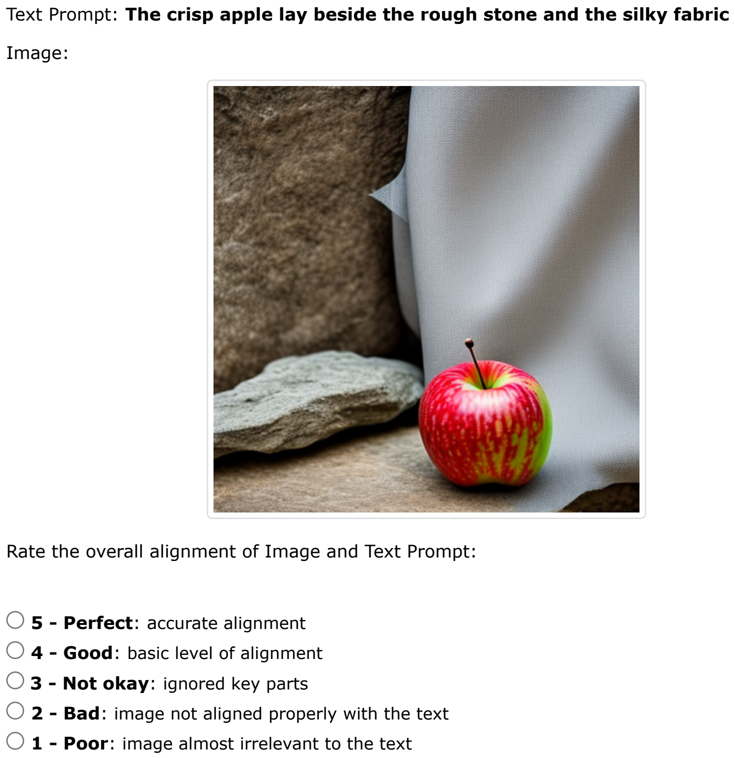
Text Prompt: **A boat is sailing on a lake**

Image:

Rate the matching degree of objects' relationship between the Image and Text Prompt:

○ **5 - Perfect**: accurate alignment
○ **4 - Good**: basic level of alignment
○ **3 - Not okay**: action relationship not correct.
○ **2 - Bad**: image not aligned properly with the text
○ **1 - Poor**: image almost irrelevant to the text

Figure 5: AMT Interface for the image-text alignment evaluation on non-spatial relationships.



Text Prompt: **The crisp apple lay beside the rough stone and the silky fabric**

Image:

Rate the overall alignment of Image and Text Prompt:

○ **5 - Perfect**: accurate alignment
○ **4 - Good**: basic level of alignment
○ **3 - Not okay**: ignored key parts
○ **2 - Bad**: image not aligned properly with the text
○ **1 - Poor**: image almost irrelevant to the text

Figure 6: AMT Interface for the image-text alignment evaluation on complex compositions.

D Additional Results

D.1 Quantitative Results of Seen and Unseen Splits

We provide the seen and unseen splits for the test set, where the unseen set consists of attribute-object pairs that do not appear in the training set. The unseen split tends to include more uncommon attribute-object combinations than seen split. The performance comparison of seen and unseen splits for attribute binding is shown in Table 6. Our observations reveal that our model exhibits slightly lower performance on the unseen set than the seen set.

Table 6: Performances of our model on attribute binding (color, shape, and texture) for seen and unseen sets.

Metric	Color		Shape		Texture	
	Seen	unseen	Seen	unseen	Seen	unseen
CLIP	**0.3422**	0.3283	0.2926	**0.3068**	**0.3240**	0.3219
B-CLIP	**0.7716**	0.7612	**0.7425**	0.6752	**0.7569**	0.6809
B-VQA	**0.7192**	0.5426	**0.5500**	0.3356	**0.7647**	0.3567
mGPT	**0.6626**	0.6780	**0.6381**	0.6307	**0.6773**	0.6580
mGPT-CoT	**0.8082**	0.8038	**0.7510**	0.6888	**0.8453**	0.7412

D.2 MiniGPT-4 Evaluation without Chain-of-Thought

Table 7 shows the additional results of benchmarking on T2I-CompBench of 6 models with MiniGPT-4 without Chain-of-Thought. Results indicate that MiniGPT-4 evaluation without Chain-of-Thought does not strictly align with human evaluation results.

Table 7: mGPT benchmarking on 6 sub-categories in T2I-CompBench.

Model	**Color**		**Shape**		**Texture**		**Spatial**		**Non-spatial**		**Complex**	
	mGPT	Human	mGPT	Human	mGPT	Human	mGPT	Human	mGPT	Human	mGPT	Human
Stable v1-4 [13]	0.6238	0.6533	0.6130	0.6160	0.6247	0.7227	0.8524	0.3813	0.8507	0.9653	0.8752	0.8067
Stable v2 [13]	0.6476	0.7747	0.6154	0.6587	0.6339	0.7827	0.8572	0.3467	0.8644	0.9827	0.8775	0.8480
Composable v2 [14]	0.6412	0.6187	0.6153	0.5133	0.6030	0.6333	0.8504	0.3080	0.8806	0.8120	0.8858	0.7520
Structured v2 [15]	0.6511	0.7867	0.6198	0.6413	0.6439	0.7760	0.8591	0.3467	0.8607	0.9773	0.8732	0.8333
Attn-Exct v2 [16]	**0.6683**	0.8240	0.6175	0.6360	0.6482	0.8400	0.8536	0.4027	0.8684	0.9533	0.8725	0.8573
GORS-unbiased (ours)	0.6668	0.8253	0.6399	0.6573	0.6389	0.8413	**0.8675**	0.4467	0.8845	0.9534	0.8876	0.8654
GORS (ours)	0.6677	**0.8320**	**0.6356**	**0.7040**	**0.6709**	**0.8573**	0.8584	**0.4560**	**0.8863**	**0.9853**	**0.8892**	**0.8680**

D.3 Reward models to Select Samples for GORS-unbiased

To avoid the bias from selecting samples by evaluation metrics as reward, we introduce new reward models which are different from our proposed evaluation metrics. Specifically, we adopt Grounded-SAM [17] as the reward model for the attribute binding category. We extract the segmentation masks of attributes and their associated nouns separately with Grounded-SAM, and use the Intersection-over-Union (IoU) between the attribute masks and the noun masks together with the grounding mask confidence to represent the attribute binding performance. We apply GLIP-based [18] selection method for spatial relationships. For non-spatial relationships, we adopt BLIP [1] to generate image captions and CLIP [8, 7] to measure the text-text similarity between the generated captions and the input text prompts. For complex compositions, we integrate the 3 aforementioned reward models as the total reward. Those sample selection models are different from the models used as evaluation metrics. The models trained with the new reward models are denoted as GORS-unbiased.

Table 8: Performances of our model on complex compositons on the 3-in-1 metric

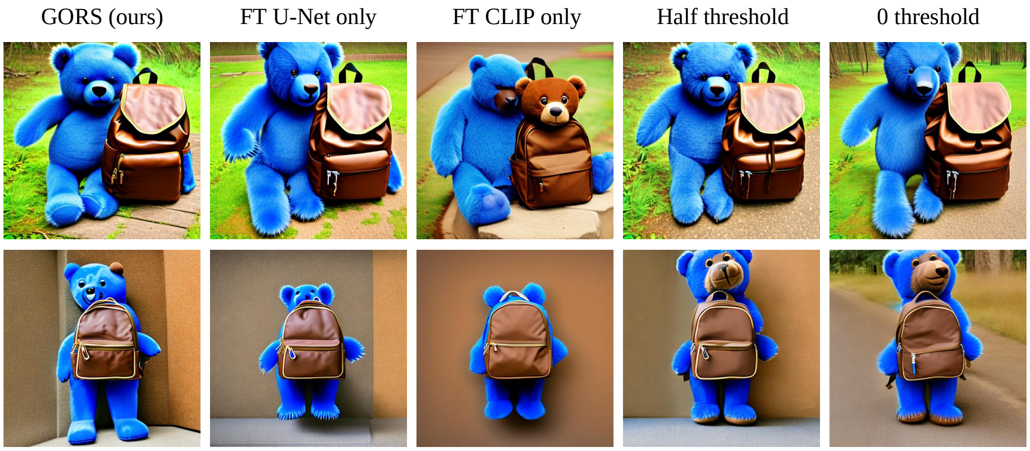
ours (350)	ours (700)	ours (1050)	ours (1400)
0.3299	0.3328	0.3371	**0.3504**

D.4 Scalability of our proposed approach

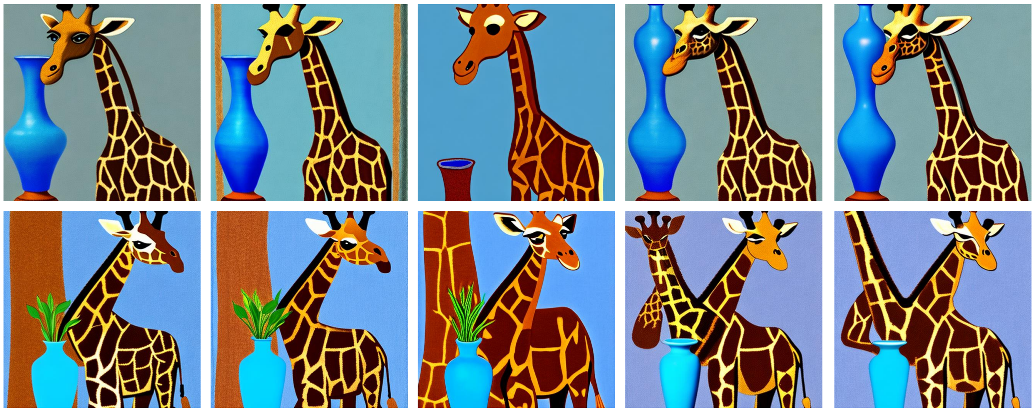
To demonstrate the scalability of our proposed approach, we introduce additional 700 prompts of complex compositions to form an extended training set of 1,400 complex prompts. The new prompts are generated with the same methodology as described in the appendix B and they are accessible through this link. We conduct 4 experiments to train the models with different training set sizes, i.e., 350 prompts, 700 prompts, 1050 prompts, and 1400 prompts. The results in Table 8 show the performance of our model grows with the increase of the training set sizes. The results indicate the potential to achieve better performance by scaling up the training set.

D.5 Qualitative Results of Ablation Study

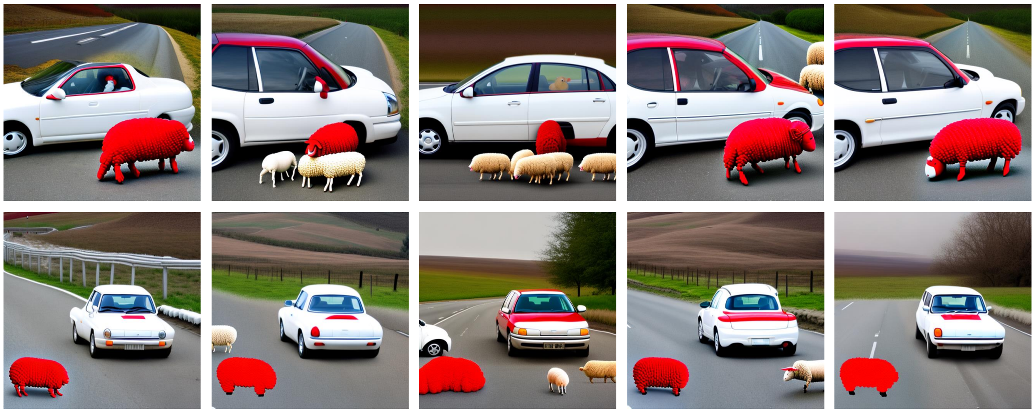
We show the qualitative results of the variants in ablation study in Figure 7. When only CLIP is fine-tuned with LoRA, the generated images do not bind attributes to correct objects (for example, Figure 7 Row. 3 Col. 3 and Row. 6 Col. 3). Noticeable improvements are observed in the generated images when U-Net is fine-tuned by LoRA, particularly when both CLIP and U-Net are finetuned together. Furthermore, we delve into the effect of the threshold for selecting images aligned with text prompts for fine-tuning. A higher threshold value enables the selection of images that are highly aligned with text prompts for finetuning, ensuring that only well-aligned examples are incorporated into the finetuning process. In contrast, a lower threshold leads to the inclusion of misaligned images during finetuning, which can degrade the compositional ability of the finetuned text-to-image models (for example, Figure 7 last two columns in Row. 2).



A brown backpack and a blue bear



A brown giraffe and a blue vase



A white car and a red sheep

Figure 7: Qualitative comparison of ablation study on fine-tuning strategy and threshold.

D.6 Qualitative Results and Comparison with Prior Work

Additional results and comparisons are shown in Figure 8 and Figure 9

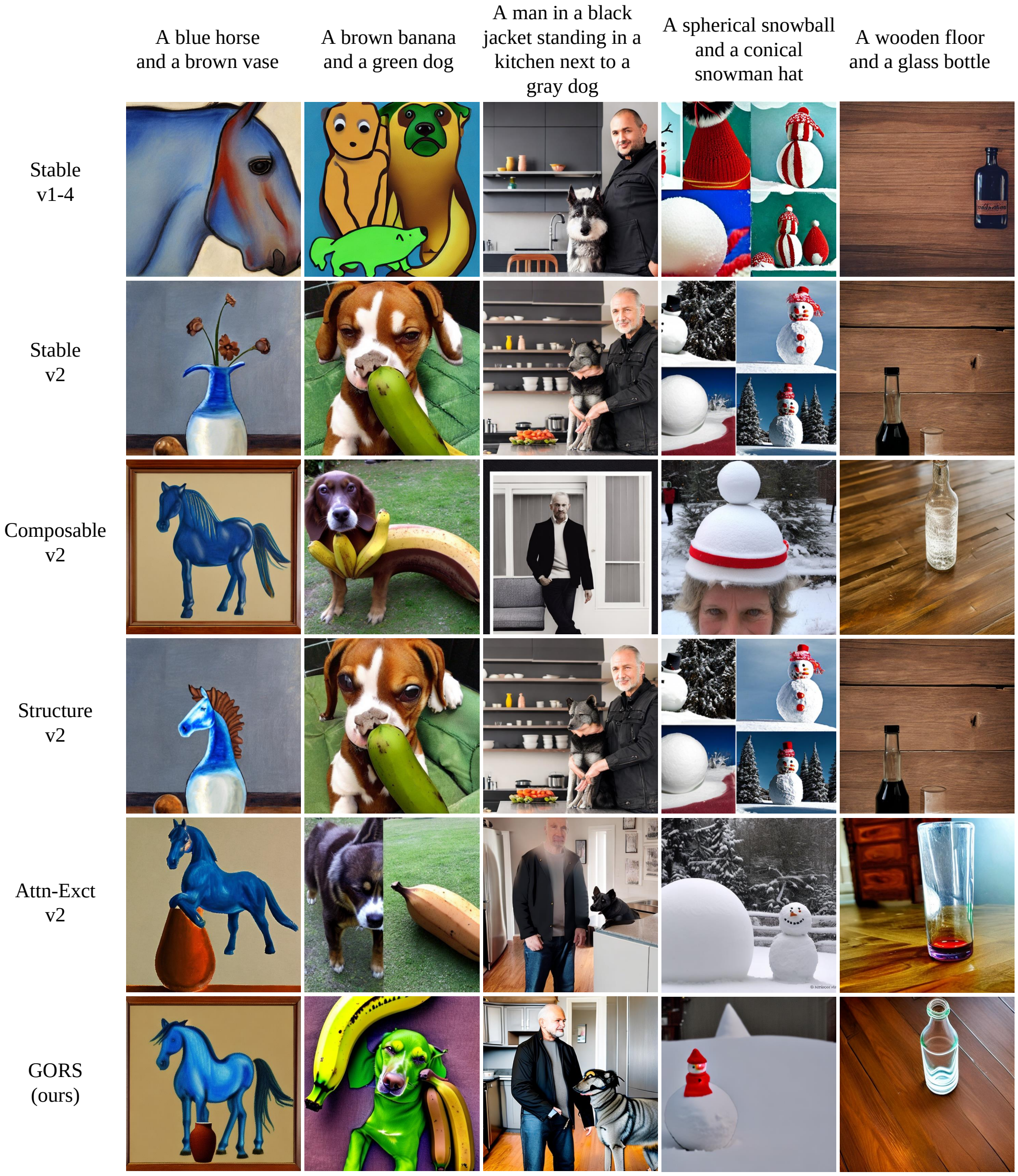

A blue horse and a brown vase
A brown banana and a green dog
A man in a black jacket standing in a kitchen next to a gray dog
A spherical snowball and a conical snowman hat
A wooden floor and a glass bottle
Stable v1-4
Stable v2
Composable v2
Structure v2
Attn-Exct v2
GORS (ours)

Figure 8: Qualitative comparison between our approach and previous methods.

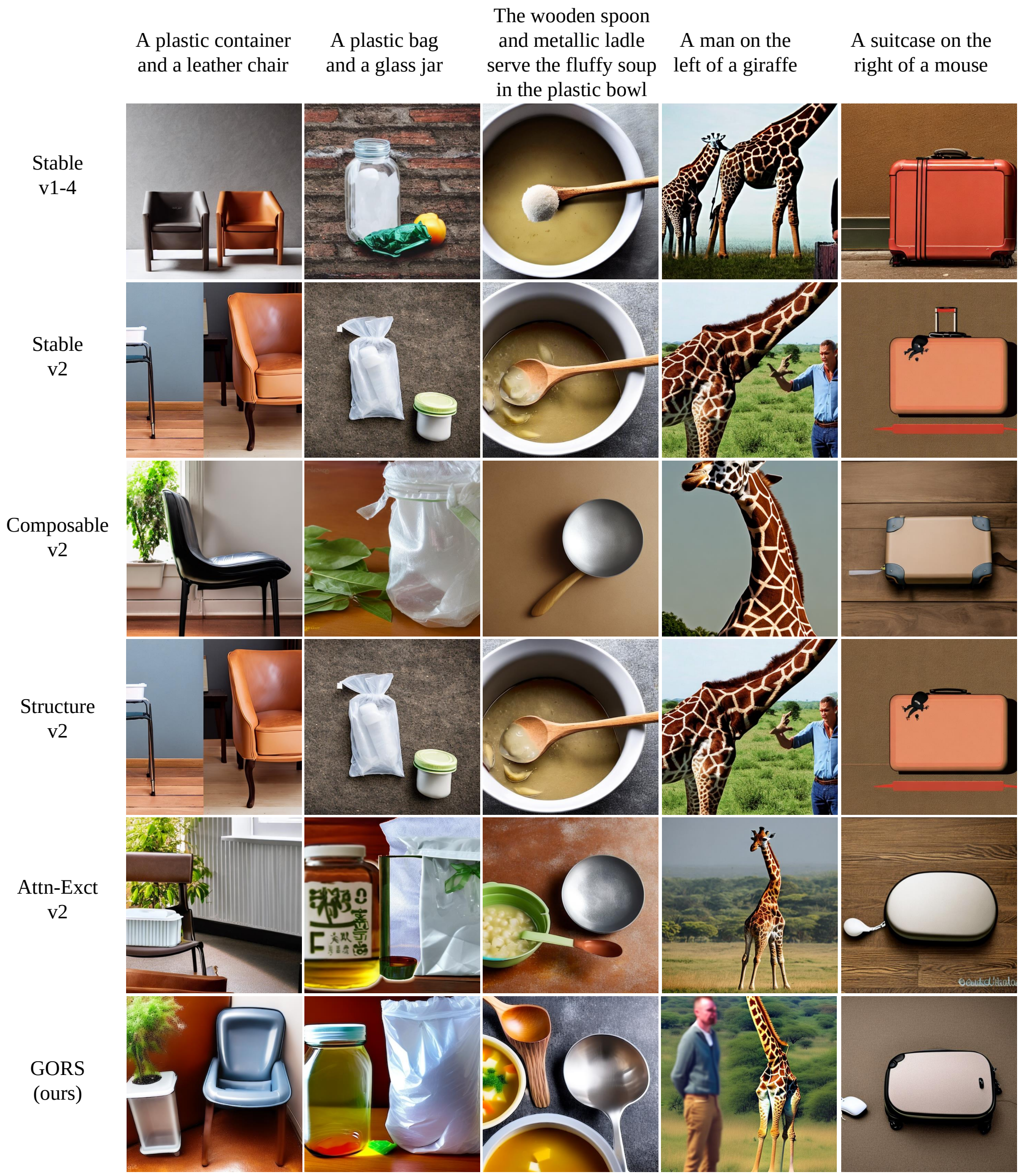
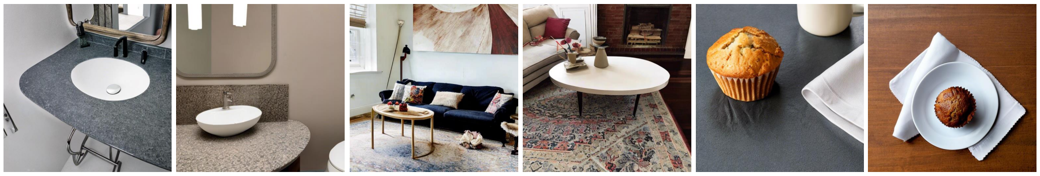

A plastic container and a leather chair
A plastic bag and a glass jar
The wooden spoon and metallic ladle serve the fluffy soup in the plastic bowl
A man on the left of a giraffe
A suitcase on the right of a mouse
Stable v1-4
Stable v2
Composable v2
Structure v2
Attn-Exct v2
GORS (ours)

Figure 9: Qualitative comparison between our approach and previous methods.



An ***oval sink*** and a ***rectangular mirror***

An ***oval coffee table*** and a ***rectangular rug***

A ***round muffin*** and a ***square napkin***

Figure 10: Failure cases of the evaluation metric BLIP-VQA.

E Limitation and Potential Negative Social Impacts

One limitation of our work is the absence of a unified metric for all forms of compositionality. Future research can explore the potential of multimodal LLM to develop a unified metric. Our proposed evaluation metrics are not perfect. As shown by the failure cases in Fig. 10, BLIP-VQA may fail in challenging cases, for example, the objects' shapes are not fully visible in the image, shape's description is uncommon or the objects are not easy to recognize. The UniDet-based evaluation metric is limited to evaluating 2D spatial relationships and we leave 3D spatial relationships for future study. Researchers need to be aware of the potential negative social impact from the abuse of text-to-image models and the biases of hallucinations from image generators as well as pre-trained multimodal models and multimodal LLMs. Future research should exercise caution when working with generated images and LLM-generated content and devise appropriate prompts to mitigate the impact of hallucinations and bias in those models.