Is Your Text-to-Image Model Robust to Caption Noise?

Supplementary Material

6. Can The Training Process Effectively Reduce Discrepancies Between Various Captions?

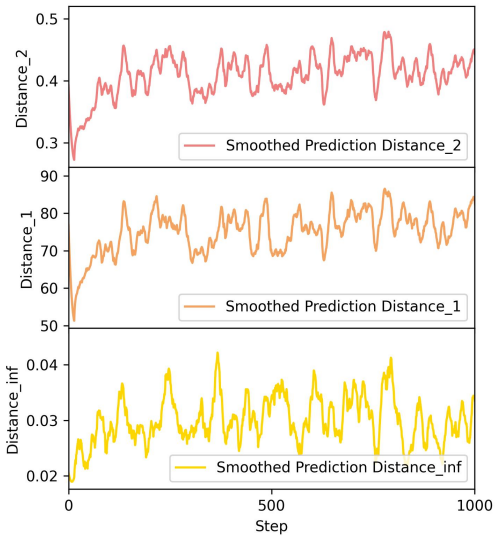


Figure 6. Prediction distance. During finetuning, with a fixed seed, we calculate the L_1, L_2 and L_∞ distance between the prediction with different captions. All of the three distances first increase and then plateau, indicating that the discrepancies between distinct captions are not reconciled through the fine-tuning process.

7. Caption Hallucination

We present examples of hallucinations in captions in Fig. 7, accompanied by several key observations regarding caption hallucinations:

1. **Variation in Hallucination Types Across Models**: Different captioning models exhibit distinct tendencies in the types of hallucinations they generate. For instance, LLaVA-v1.6 demonstrates a higher prevalence of hallucinations related to color and specific features, while Share-Captioner shows greater susceptibility to hallucinations involving color and spatial relationships.
2. **Proportion of Hallucinated Content**: Hallucinated content constitutes only a minor fraction of the overall caption. Current captioning models and vision-language models (VLMs) generally achieve high accuracy, producing predominantly correct captions with hallucinations typically confined to specific attributes.
3. **Influence of Text Exposure Bias in Language Models**: Hallucinated content can be attributed, in part, to text exposure bias in large language models (LLMs).[3] For example, during LLM pretraining, "envelopes" are frequently associated with "mailing", which leads to the hallucination of the phrase "which are likely for mailing purposes" in the second row of the example.
4. **Hallucination Types**: Hallucinations in captions can be categorized into four primary types: color, spatial relationships, quantity, and specific features. To evaluate the quality of generated captions along these dimensions, we have developed a specialized benchmark, InstructBench, designed to assess performance with respect to these attributes.

8. Prompts for Hallucination Rate

As described in Section 4.2, we prompt GPT-4o for extracting hallucinated contents and calculate the hallucination rate to evaluate the quality of image captions. We design a two-round prompt for objects extraction and hallucination justification separately. Details can be found in Figure 8 and Figure 9.

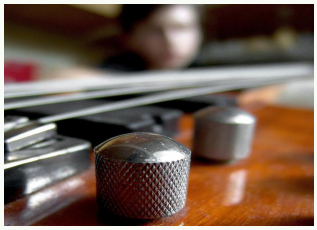
We first use the prompt in Figure 8 to extract visual components/object from a given detailed image caption. Then, we provide GPT-4o with both the image and the prompt in Figure 9 to justify the existence of each extracted visual component/object.

9. Visualization

Apart from Fig.5, we also provide more visualizations in Fig. 10. Here are some observations.

1. The proposed method demonstrates a superior ability to adhere to the provided instructions, as evidenced by its higher performance on the InstructBench evaluation benchmark. For instance, in the visualization corresponding to the prompt "a mix of flowers featuring bright red and bright yellow," the proposed method generates images strictly confined to the specified colors of bright red and bright yellow. In contrast, alternative methods may inaccurately include flowers of intermediate hues, such as orange. Similarly, for prompts like "a folding chair by the window," baseline methods frequently fail to satisfy both constraints, either producing an image with a chair that is not folding or omitting the detail of it being positioned "by the window."
2. the proposed method exhibits varying degrees of improvement across different attributes. For attributes such

[3] Text exposure bias refers to the discrepancy between training and testing conditions caused by the reliance on teacher forcing during maximum likelihood estimation.



The image shows a close-up of a person's hands playing a stringed musical instrument, likely a guitar or a similar stringed instrument. The focus is on the hands and the instrument, with the background being out of focus. The person appears to be seated and is wearing a dark-colored shirt. The instrument has a wooden body with a black pickguard and a metal neck. There are two silver-colored pickups mounted on the neck, and the person is pressing down on the strings with their fingers. The image is taken from a low angle, looking up at the hands and the instrument. There is no visible text in the image.

The image shows a box of KNOX Gelatine, which is a type of food product. The box is predominantly white with orange and black text and graphics. The brand name \"KNOX\" is prominently displayed at the top of the box in black text. Below the brand name, the product name \"Gelatine\" is written in large, bold, orange letters. The box also features a graphic of a spoon and a piece of cake, suggesting the product's use in baking and desserts. The box indicates that the gelatine is \"UNLEVELLED\" and \"ORIGINAL,\" and it is \"SATISFACTION GUARANTEED.\" The box also mentions that it contains \"4\" envelopes, which are likely for mailing purposes. The background of the image is a patterned surface with a mix of red and yellow colors.

In the image, there are three individuals standing in front of a gate. The gate, which is the central focus of the image, is adorned with a wrought iron design. \n\nOn the left, there's a person dressed in a black and white checkered shirt paired with blue jeans. This individual is holding a black purse in their hand. \n\nIn the middle, there's another person who is wearing a black jacket and blue jeans. This person is also holding a black purse. \n\nLastly, on the right, there's a person wearing a green jacket and orange shorts. This person is holding a white purse. \n\nAll three individuals are looking directly at the camera, creating a sense of engagement with the viewer. The relative positions of the individuals and their attire suggest a casual, relaxed atmosphere. The image does not contain any discernible text. The overall composition of the image is balanced, with the three individuals and the gate serving as the main elements.

The image presents a tranquil scene of a two-story house nestled amidst nature. The house, painted in a pristine shade of white, stands out against the lush greenery that surrounds it. Its gray roof and chimney add a touch of contrast to the otherwise monochromatic structure.\n\nThe house is encircled by a well-manicured lawn, a vibrant carpet of green that stretches out in all directions. The lawn is dotted with several bushes and trees, their leaves a mix of different shades of green, adding depth and texture to the scene.\n\nThe sky above is a clear blue, devoid of any clouds, suggesting a calm, sunny day. The absence of any text or human activity in the image further enhances the serene and peaceful atmosphere of the scene. The relative positions of the objects - the house, the lawn, the bushes, and the trees - create a harmonious balance, making the image a picturesque representation of a quiet, suburban setting.

Figure 7. Examples of the caption hallucination in the dataset. The words highlighted in yellow denotes the hallucination which unaligns with the image. The top two rows are captioned using LLaVA-7B-v1.6, and the bottom two rows are captioned using Share-Captioner.

as color fidelity and spatial positioning, the enhancements are particularly pronounced. However, for attributes related to quantity, such as ensuring a specific number of objects, the improvements are comparatively less significant. This suggests that the method excels in some aspects of semantic precision but still faces challenges in others.

Failure Cases Analysis. We also present several failure cases in Fig. 11, where the proposed model struggles to generate the intended features. And here are two observations.

1. These issues are particularly pronounced in generating quantitative attributes. This limitation likely stems from an imbalance in the training corpus, where quantitative terms appear significantly less frequently than color-related terms. For instance, in a randomly sampled subset of 10,000 training examples, the term 'four' appears 482 times, and 'five' appears 89 times, whereas the term 'red' occurs 11147 times, and 'white' occurs 10619 times.
2. We observe that the generated images often exhibit watermarks, as at the bottom of the fourth images in the ex-

Category	Keywords	Generation Prompt	Evaluation Prompt
Spatial	Left, right, vertical, horizontal, inside, outside, middle, corner, *etc*.	The dog's head is turned slightly to the right.	Is the dog's head turned slightly to the right?
Color	Yellow, orange, purple, pink, brown, light blue, navy blue, *etc*.	A kitchen with floor of a light blue color.	Does the kitchen have a floor of a light blue color?
Quantity	Number of objects/living creatures, from one to six.	The image shows four individuals posing for a photograph.	Are there four individuals posing for a photograph?
Features	Droopy, locked, unlocked, open, closed, full, empty, invisible, *etc*.	A cat with eyes open.	Are the cat's eyes open in this image?

Table 4. Category and examples of InstructBench. Each category comprises 200 generation-evaluation pairs. These categories align with the four primary types of hallucinations typically observed in captioning tasks.

```
# to extract visual components (objects) from a given CAPTION
You are an expert in extracting visual components from image descriptions.
For the given detailed description, you need to list all the visual components in the description, including objects,
texture, environment, etc.
Do not count a single object twice. Do not count any conjucture.
Do not include the atmosphere as visual components.
Do not include things that are not visible as visual components.
Do not include motions that have not been done as visual components.

Here are three samples:

Description: The image shows a spoon is filled with a yellow substance, possibly honey or mustard, and it is being
     lifted from a bowl. The spoon is held by someone who is not visible in the frame. The background features a
     wooden table, which adds to the overall homey atmosphere of the scene. The focus is on the spoon and its contents
     , emphasizing the texture and color of the substance being scooped up.
You should output visual components in this format: ["a spoon", "a yellow substance", "a bowl", "a wooden table"]

Description: The image depicts a street scene with a focus on a building and a car. The building appears to be a two-
     story structure with a flat roof, possibly a commercial or residential building. There are several cars parked or
      moving along the street in front of the building. The street itself is lined with trees and a sidewalk, and
     there are traffic lights visible at the intersection. The sky is partly cloudy, suggesting a fair weather
     condition. The image is in color and has a standard resolution. There are no visible texts or distinctive brands
     in the image.
You should output visual components in this format: ["a street scene", "a car", "a two-story structure building with a
      flat roof", "several cars parked or moving along the street", "the street lined with trees and a sidewalk", "
     traffic lights at the intersection", "partly cloudy sky"]

Description: The image shows a snake resting in a curved container. The container appears to be made of a material
     that could be a ceramic or a similar type of enclosure. The snake has a patterned body with shades of brown,
     black, and yellow. It is coiled up and seems to be in a relaxed state, possibly sleeping or resting. The
     container is placed on a bed of straw-like material, which provides a naturalistic environment for the snake. The
      background is not clearly visible due to the close-up nature of the photograph. There are no visible texts or
     markings on the image.
You should output visual components in this format: ["a snake resting in a curved container", "a ceramic container", "
     patterned snake body with shades of brown, black, and yellow", "a snake coiled up", "a container placed on a bed
     of straw-like material"]

Description: {CAPTION}
Can you output the visual components as json following the above format?
```

Figure 8. First prompt used to extract visual components/objects from a given detailed image caption. We provide three in-context-learning examples to instruct GPT-4o for object extraction. The output will follow the json format for parsing.

amples, which can confuse the generation model. This artifact arises from the prevalence of watermarked images in the training dataset.

10. Implementation Details

In Fig. 2, we use GPT4o-mini to label a subset of hallucinated content, we use the following prompts in the inputs to generate the hallucinated words in the caption.

Hallucinated word generation prompt:
In the following caption of the image, which words are **not** faithfully describing the image? List 1)the words and their positions, and a revised version of the caption based on the original caption, in a json format.

```
# ask GPT-4o to determine the existence of each extracted object
You are a smart expert in evaluating visual components in images.
Here is a list of visual components that are possible to exist in the provided image: {}
In the list, visual components are split by the punctuation comma. Please consider them separately.
Ask yourself: Can you see [component] in the image?
The [component] can be substituted by each element in the list of visual components.
Your answer should be a json of a dictionary of 1/0 answers.
1 means yes, 0 means no, similar to {{"component_i": 1/0, ...}}
```

Figure 9. Second prompt used to determine the existence of each extracted object from the previous object extraction step. We provide GPT-4o with both input image and the prompt including extracted object to determinine if each extracted object exists in the given image.

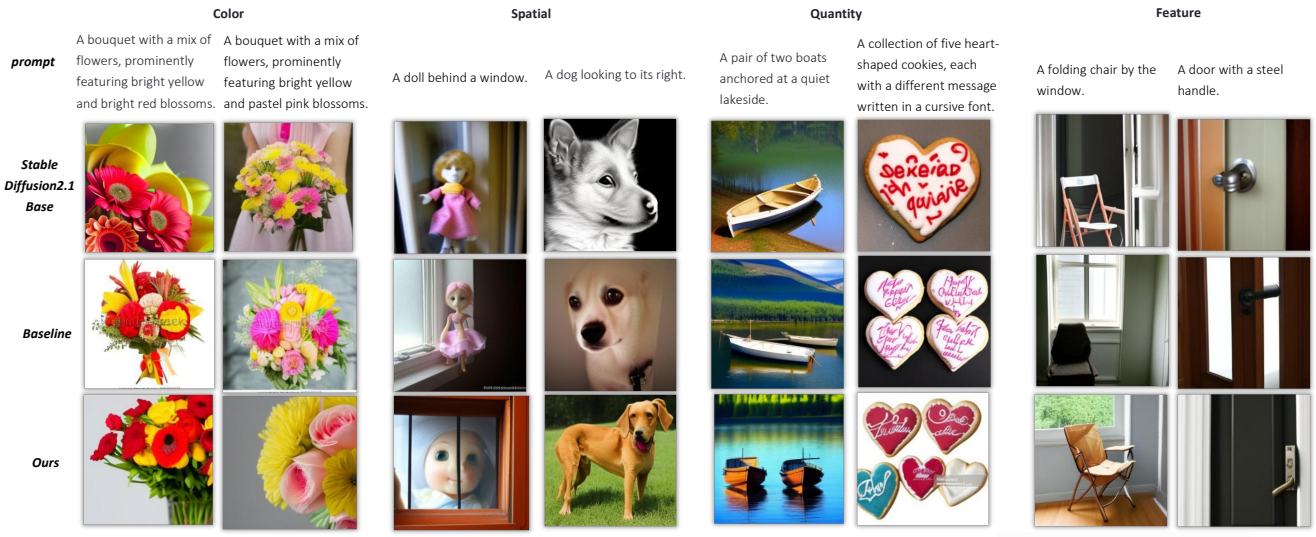


Figure 10. More examples of the generated images on InstructBench. The top row represents outputs from the original Stable Diffusion 2.1-base model. The middle row is our base model, finetuned on the caption dataset without specific mitigation strategies for caption hallucination. The last row showcases our model trained with the proposed robust training framework. We split the generated images into four dimension: color, spatial, quantity, and feature. We observe that our method better follows the text prompts.

{word1: position1, word2: position2,..., caption: revised_caption}.

When using VLM to compute the confidence score in Eq. (2), we use the following prompts in the inputs of the VLM before the input caption and image.

Prompt for later calculating VLM confidence score:

You are a powerful image captioner. Provide a detailed description of the image.

Instead of describing the imaginary content, only describing the content one can determine confidently from the image. Do not describe the contents by itemizing them in list form. Minimize aesthetic descriptions as much as possible.

A quartet of four boats anchored at a quiet lakeside.

A quartet of four painters creating art in a bright studio.

A group of five birds perched on a tree branch.

A band is playing music, with their name displayed in white text at the bottom of the image.

Figure 11. Failure cases of generated images on InstructBench. The failure cases are particularly pronounced in generating quantitative attributes. And a main reason may be the deficiency in training dataset. For instance, in a randomly sampled subset of 10,000 training examples, the term 'four' appears 482 times, and 'five' appears 89 times, whereas the term 'red' occurs 11147 times, and 'white' occurs 10619 times.