PlanGen: Towards Unified Layout Planning and Image Generation in Auto-Regressive Vision Language Models

Supplementary Material

A. Inference Details

We used classifier-free guidance during the inference process of image generation, and the guidance scale is 5. We only apply negative layout gudance when performing layout-guided image manipulation including object deletion. To speed up the inference, we use kv-cache. It takes about 17 seconds to generate 8 images in one batch on a single A100 GPU.

B. Experimental Details

In-context Prompt for Baselines in Layout Planning. Qwen-2.5-7b-instruct and Llama-3.1-8b-instruct themselves do not naturally support the layout planning task. We leverage LLMs' in-context learning capabilities to generate layouts from global captions following LayoutGPT, and the in-context prompts are as follows:

```
You are an intelligent bounding box
    generator. I will provide you with a
     caption for a photo, image, or
    painting. Your task is to generate
    the bounding boxes for the objects
    mentioned in the caption, along with
     a background prompt describing the
    scene. The images are of size 512
    x512. The top-left corner has
    coordinate [0, 0]. The bottom-right
    corner has coordinnate [512, 512].
    The bounding boxes should not
    overlap or go beyond the image
    boundaries. Each bounding box should
     be in the format of (object name, [
    top-left x coordinate, top-left y
    coordinate, box width, box height])
    and should not include more than one
     object. Do not put objects that are
     already provided in the bounding
    boxes into the background prompt. Do
     not include non-existing or
    excluded objects in the background
    prompt. Use "A realistic scene" as
    the background prompt if no
    background is given in the prompt.
    If needed, you can make reasonable
    guesses. Please refer to the example
     below for the desired format.

input: A realistic image of landscape
   scene depicting a green car parking
   on the left of a blue truck, with a
   red air balloon and a bird in the
   sky
you need output: [('a green car', [21,
   281, 211, 159]), ('a blue truck',
   [269, 283, 209, 160]), ('a red air
   balloon', [66, 8, 145, 135]), ('a
   bird', [296, 42, 143, 100])]

input: A realistic top-down view of a
   wooden table with two apples on it
you need output: [('a wooden table',
   [20, 148, 472, 216]), ('an apple',
   [150, 226, 100, 100]), ('an apple',
   [280, 226, 100, 100])]

input: An oil painting of a pink
   dolphin jumping on the left of a
   steam boat on the sea
you need output: [('a steam boat',
   [232, 225, 257, 149]), ('a jumping
   pink dolphin', [21, 249, 189, 123])]

The input for you to process is: {}
```

Baselines Details on Image Layout Understanding. For Grounding-DINO, we perform grounding detection by giving the image and the global caption of the image. For two other LLM-based baselines, i.e. CogVLM-grounding and Qwen-VL-Chat, we take prompts suitable for the corresponding models to help the model to output accurate detection results. Specifically, for CogVLM-grounding, we give the image and ask "*Can you provide a description of the image and include the coordinates [[x0,y0,x1,y1]] for each mentioned object?*" following its formal demo. For Qwen-VL-Chat, we apply two rounds of questions. First, we give the image and ask the model "*What objects are in the image?*". Then, after the model answers this question, we ask the model "*Box out the positions of these objects in the figure*". We find that for Qwen-VL-Chat, the effect of such two-round Q&A will be much better than a single question.

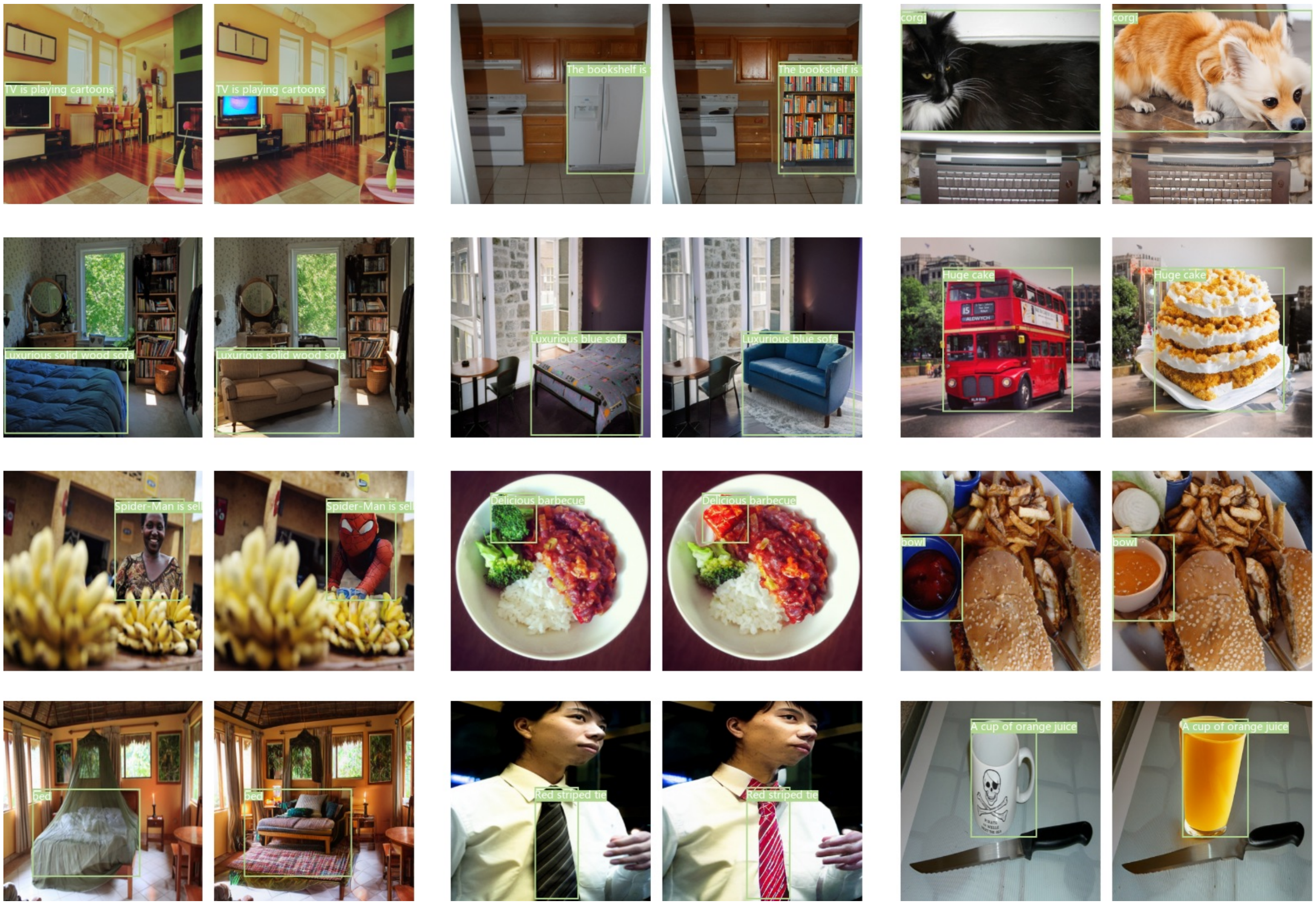


Figure 1. More examples for Layout-guided Image Manipulation. The contents to be edited are drawn in the form of bounding boxes on the original images and the edited images for easy comparisons.

Evaluation of Image Layout Understanding. Similar to HiCo, we calculate the maximum IoU between the boxes predicted by different models and the ground truth boxes. If the maximum IoU is higher than the threshod 0.5, we caculate the clip text similarity between corresponding local descriptions. If the CLIP score is higher than 0.2, we mark it as a correct prediction. We use AR, AP, AP50 and AP75 to evaluate the performance of image layout understanding.

C. Additional Results

We show more examples of layout-guided image manipulation in Figure 1, more examples of layout-image joint generation in Figure 3, more examples of layout-to-image in Figure 4, and more results of image layout understanding in Figure 5. These rich examples show PlanGen's excellent performance on multiple related tasks.

Comparisions with Janus-Pro Baseline. We supplement comparisons with the Janus-Pro (1B) baseline across multiple tasks. PlanGen significantly outperforms Janus-Pro in layout-to-image as shown in Table 1, with superior region-wise scores and a lower FID. We observe that even with in-context examples, Janus-Pro (1B & 7B) struggles to generate reasonable layouts from captions, potentially due to its limited training. Similarly, Janus-Pro faces challenges in understanding image layouts. All these experiments show that PlanGen holds significant advantages over the Janus-Pro baseline.

Method	Spatial ↑	Color. ↑	Textual. ↑	Shape ↑	FID ↓
Janus-Pro	72.91	59.67	62.85	61.03	18.52
PlanGen	**92.21** (+19.30)	**82.69** (+23.02)	**86.53** (+23.68)	**85.36** (+24.33)	**13.91** (-4.61)

Table 1. Layout-to-image comparision with Janus-Pro baseline.

Failure cases. We also show several failure cases on the task of layout-image joint generation in Figure 2. We observed that PlanGen may experience distortion when generating human bodies, as shown in Figure 2, which is a common challenge faced by some previous autoregressive image generation models. When multiple identical objects

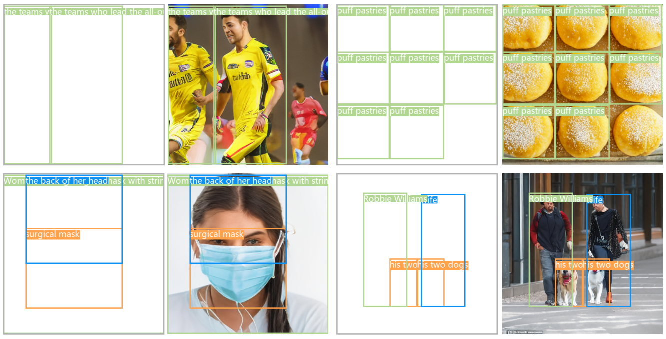


Figure 2. Failure cases.

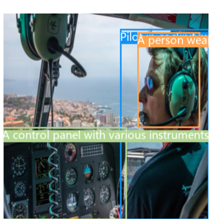
are generated, additional ones may appear, which is caused by incomplete object annotations in the training data. In the 2nd row, the the geometry of generated mask straps is slightly inappropriate. When generating dogs in smaller size, the quality declines also occur. More efficient image modeling or wider training should alleviate these problems.

Figure 3. More examples for Layout-Image Joint Generation. Global captions for layout-image joint generation are attached below the images.

Figure 4. More examples for Layout-to-Image Generation.

This is a photo showcasing a bustling street in a city, with two women walking down the street in the foreground. They are dressed in casual clothes, one wearing a black coat and the other in a black top, both carrying backpacks. The street is lined with shops and restaurants, with signs and advertisements in both Chinese and English, indicating that this is a multicultural area. The street is also filled with pedestrians and vehicles, creating a typical urban life scene. **<grounding>**<ref>Two women, one in a black coat and the other in a pink jacket with a backpack, walking on a city street.</ref><box>[374, 316, 689, 991]</box><ref>A woman with long dark hair, wearing a black jacket and carrying a striped backpack.</ref><box>[533, 319, 688, 962]</box><ref>A woman with long dark hair, wearing a black coat and carrying a blue backpack.</ref><box>[377, 320, 526, 992]</box>**</grounding>**

This is a photo showing a commercial jetliner landing. The plane is in the process of landing, with its landing gear already deployed. The fuselage of the plane is primarily white, with red and blue markings. The background is a clear blue sky and a row of green trees, with a fence in the foreground. **<grounding>**<ref>A commercial jetliner with a white fuselage, red and blue tail, and the word "ROSSIA" on its side is in flight.</ref><box>[343, 411, 799, 662]</box><ref>A dense, lush forest with a variety of green hues.</ref><box>[0, 381, 1000, 801]</box><ref>A clear, cloudless blue sky.</ref><box>[0, 2, 1000, 444]</box><ref>A green chain–link fence with barbed wire on top, set against a backdrop of dry grass and trees.</ref><box>[0, 766, 1000, 996]</box>**</grounding>**



This is a photo showing a pilot in a cockpit. The pilot is wearing a green headset, with his back to the camera, looking out the window. The cockpit is equipped with a control panel, which is filled with various instruments and switches. Through the window, you can see the cityscape outside, including buildings and the sea. The entire scene is illuminated by natural light, creating a professional and focused atmosphere. <**grounding**><ref>Pilot in cockpit, wearing headset and sunglasses, looking out at cityscape.</ref><box>[572, 85, 1000, 997]</box><ref>A person wearing a green headset with reflective sunglasses.</ref><box>[656, 100, 956, 556]</box><ref>A control panel with various instruments and switches, including a clock, indicator lights, and a digital display.</ref><box>[0, 553, 602, 997]</box></**grounding**>

Figure 5. More examples for Image Layout Understanding.