Grounded Text-to-Image Synthesis with Attention Refocusing

Quynh Phung Songwei Ge Jia-Bin Huang

University of Maryland College Park

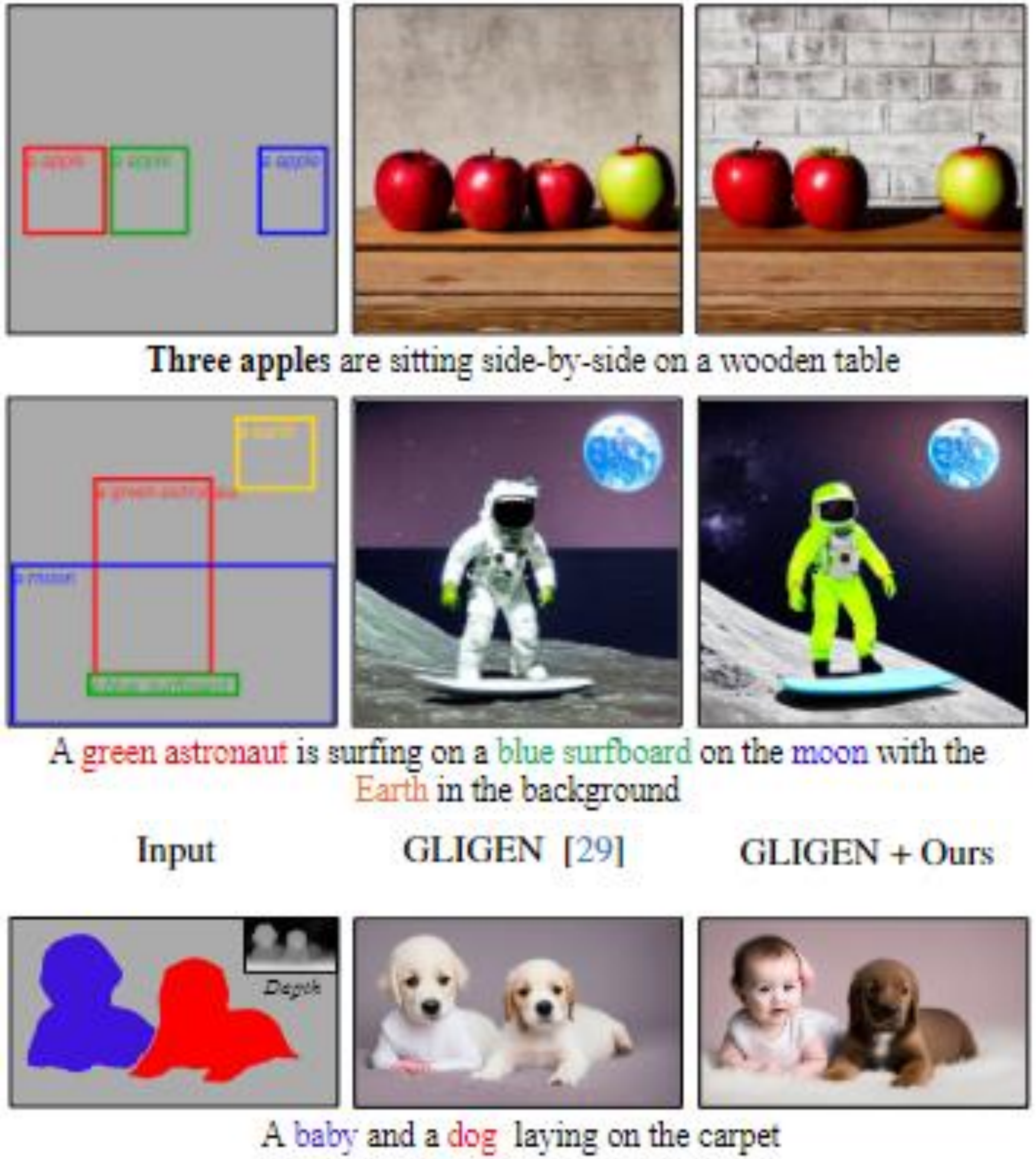
Outline

- Introduction
- Related work
- Method
- Experiments
- Limitation

Introduction

Problem Statement

- Models struggle with complex prompts involving multiple objects, attributes, and spatial relationships.
- Issues include objects being mixed, swapped, or missing due to problems in the attention layers of models like Stable Diffusion.
- Specifically, similar pixels can cause incorrect attention assignment, leading to errors like missing or blended attributes.



Proposed Solution

- The authors propose explicit spatial layouts as part of the solution.
- Two new attention-refocusing losses are introduced to improve both self- and cross-attention layers.
- These losses help ensure that attention is refocused on the correct regions, preventing mixing between different objects.

Contribution

- Novel losses that refocus attention during the sampling process, improving control over the image generation based on text prompts and layouts.
- Use of LLMs to generate layouts for better grounding in text-to-image synthesis.
- Comprehensive experiments show significant improvement over existing methods on benchmarks like DrawBench, HRS, and TIFA.

Related work

Large-scale text-to-image models

- Key Techniques:
  - Availability of large-scale text-image datasets enables training on diverse, large-scale data.
  - Development of scalable architectures: GANs, autoregressive models, diffusion models.
  - Enhanced training and inference techniques for model improvement.
- Focus of Research: Improving the controllability of generated images based on input text, using large-scale diffusion models.

Improving the controllability of text-to-image models

- Challenges: Text-to-image models often struggle with fulfilling complex prompts, leading to missing or mixed elements.
- Key Approaches:
  - Various input formats: rich text, personal images, edge maps, segmentation masks, depth maps, bounding boxes.
  - Enhancing control through improved text alignment: Attend-and-Excite method, human feedback, improved language models.
- Our Work: We focus on leveraging intermediate spatial layouts generated by large language models (LLMs) to ground image synthesis.

Layout-conditioned text-to-image synthesis

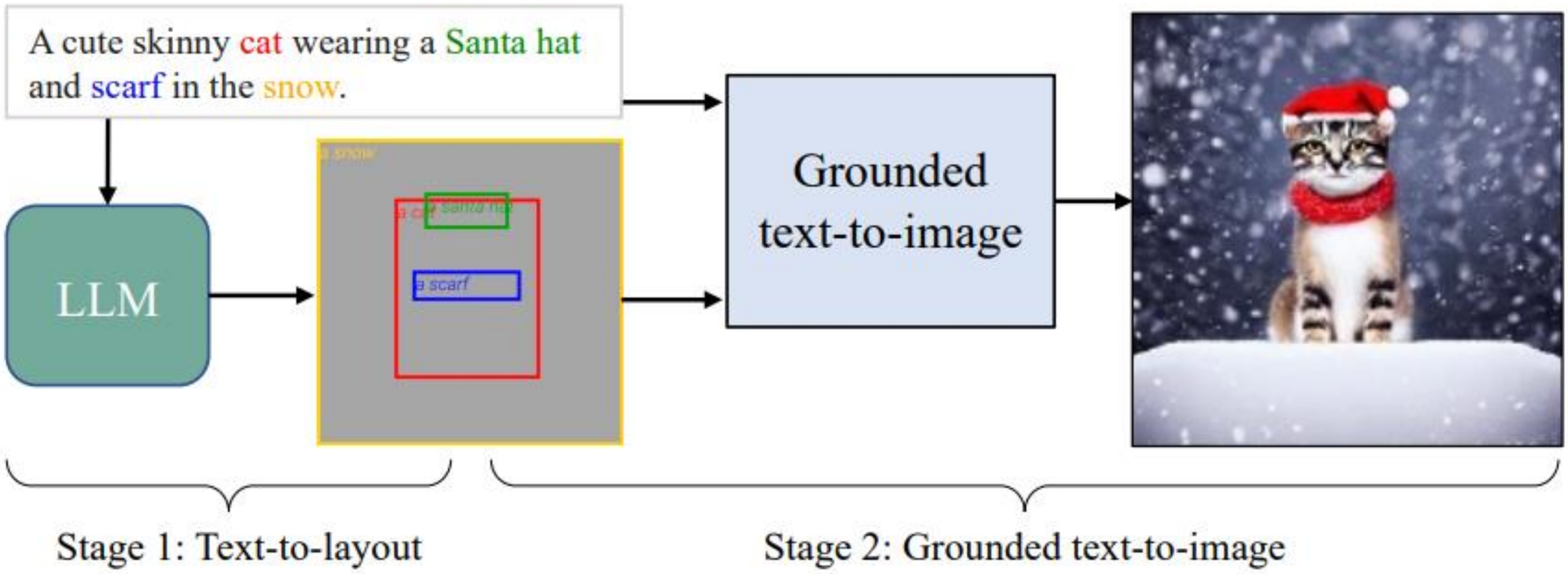
- Optimizes both cross-attention and self-attention maps, iteratively improving peak values without degrading image quality.
- Comparison:
  - Other methods like DenseDiffusion modify attention maps without optimization.
  - Our method optimizes the latent space under mask guidance, maintaining image quality.
- Result: Our attention-based guidance consistently improves performance across base models without extra training.

Layout predictions

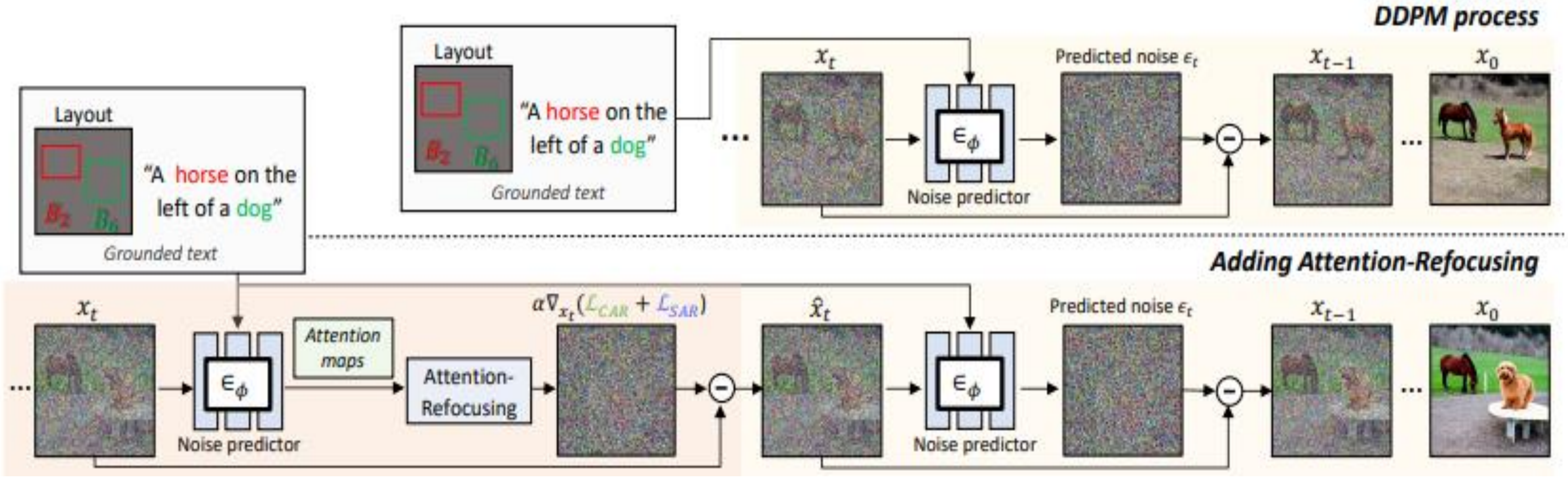
- Use of LLMs: Several works, like LLM-grounded Diffusion and LayoutGPT, use large language models (e.g., GPT-4) to generate scene layouts for text-to-image generation.
- Challenges: Current models struggle with accurately representing details such as quantity, identity, and attributes from text prompts.
- Improvement: This paper enhance controllability by using attention-based guidance, offering better alignment between text and generated images.

Method

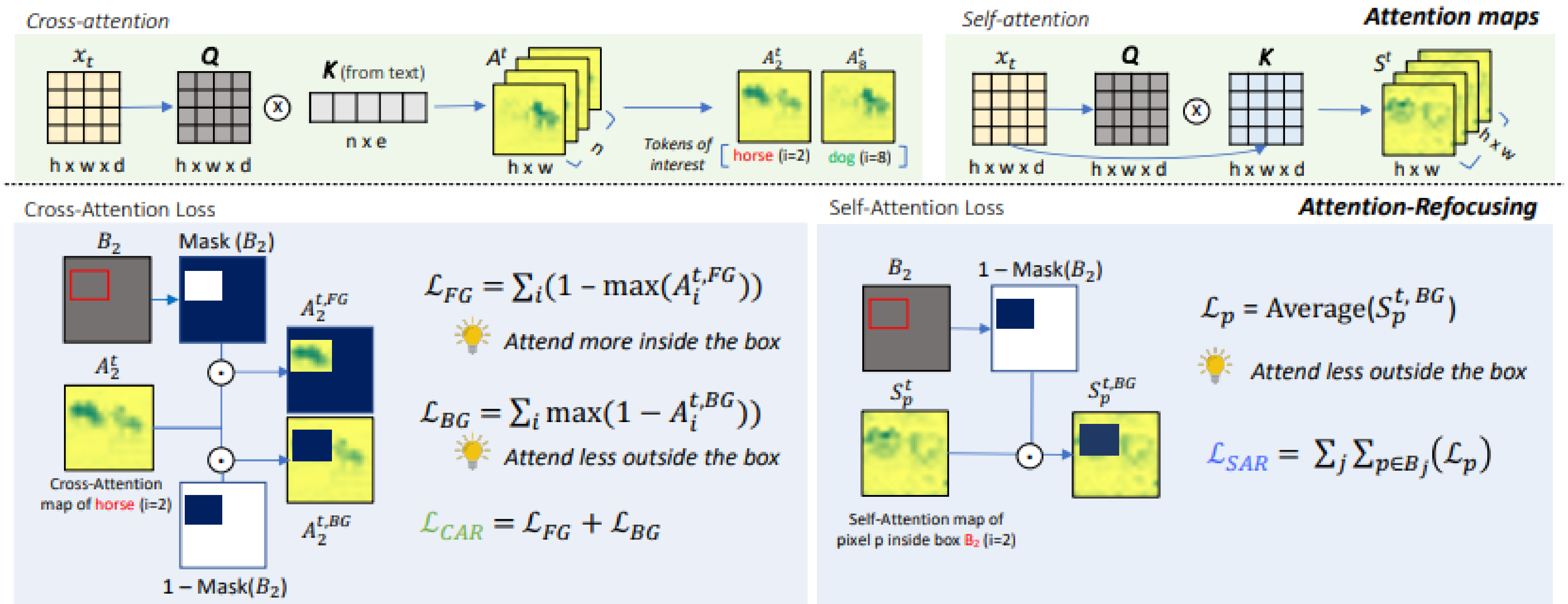
Pipeline


A cute skinny cat wearing a Santa hat and scarf in the snow.
LLM
a scarf
Grounded text-to-image
Stage 1: Text-to-layout
Stage 2: Grounded text-to-image

Attention-Refocusing framework



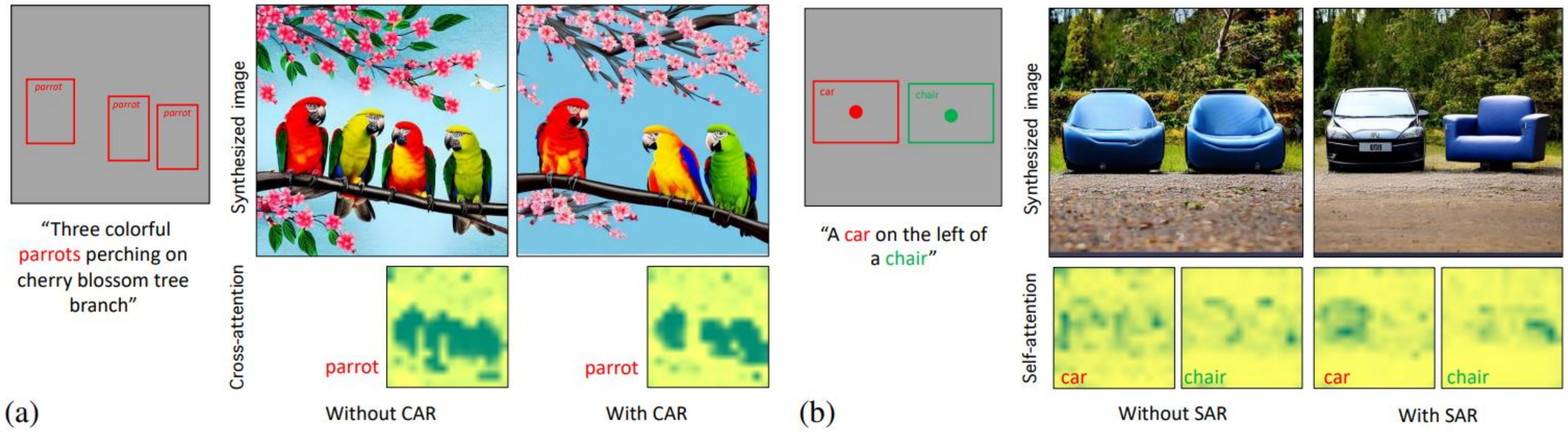
Attention-Refocusing framework



$$\mathcal{L}_{FG} = \frac{1}{q} \sum_{i \in I} (1 - \max(A_i^t \cdot \text{Mask}(B_i)))$$

$$\mathcal{L}_{BG} = \frac{1}{q} \sum_{i \in I} \max(A_i^t \cdot (1 - \text{Mask}(B_i)))$$

Grounded text-to-image generation

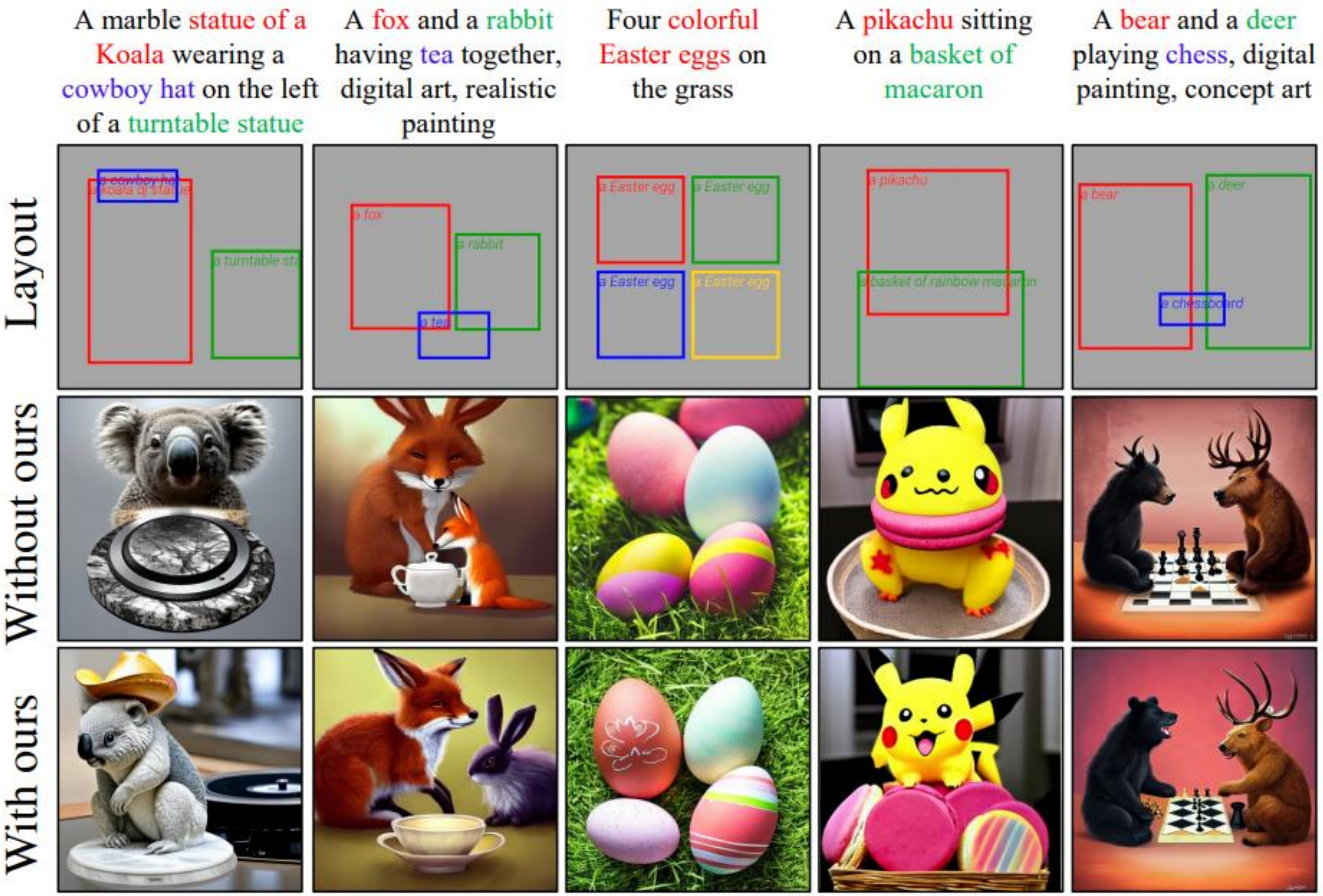


Text-to-layout prediction

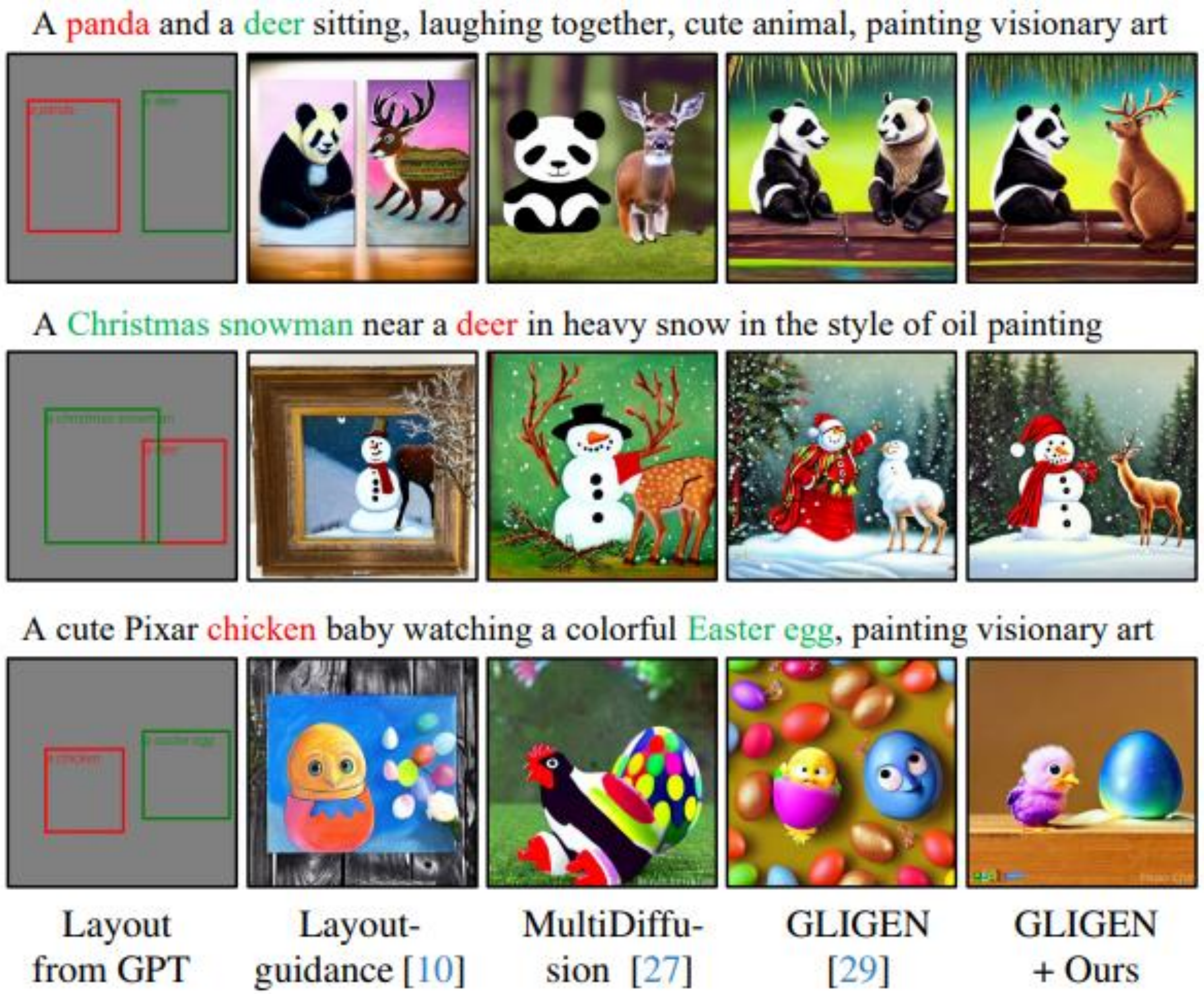
Role	Content
Instruction	System: "You are ChatGPT-4, a large language model trained by OpenAI. Your goal is to assist users by providing helpful and relevant information. In this context, you are expected to generate specific coordinate box locations for objects in a description, considering their relative sizes and positions and the number of objects. The box coordinates should be in the order (left, top, right, bottom). The size of the image is 512*512."
In-context examples	User: "Provide box coordinates for an image with a cat in the middle of a car and a chair. Make the size of the boxes as big as possible." Assistant: "cat: (245, 176, 345, 336); car: (10, 128, 230, 384); chair: (353, 224, 498, 350)" User : "Provide box coordinates for an image with three cats on the field." Assistant: "cat: (51, 82, 399, 279);cat: (288, 128, 472, 299); cat: (27, 355, 418, 494)"
User prompt	User : "Provide the Provide box coordinates for an image with" + [user prompt]

Experiments

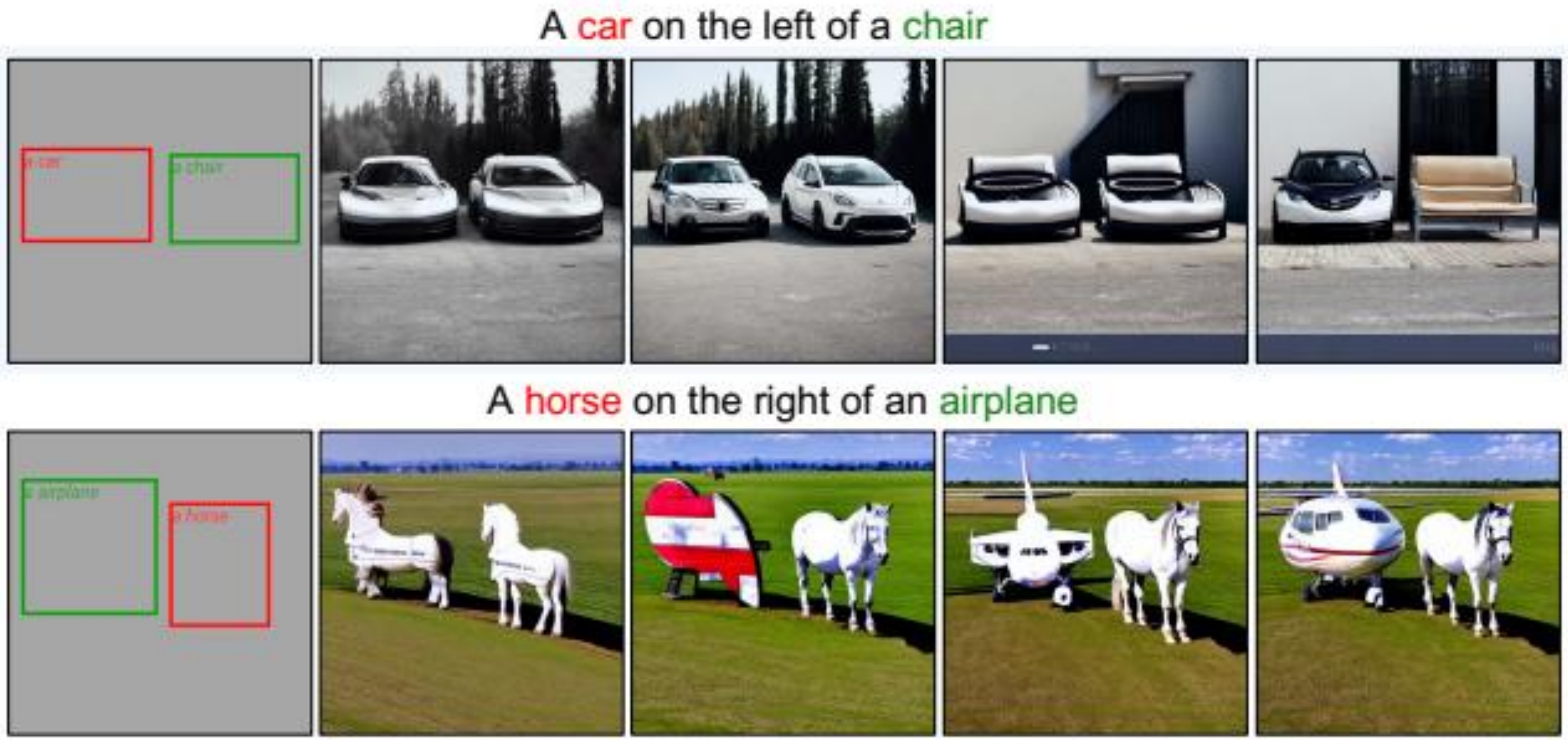
Plug & play use of our attention-based guidance



Visual comparisons on HRS benchmark



Ablation study



CAR	×	×	✓	✓
SAR	×	✓	×	✓

Ablation study

CAR	SAR	Counting			Spatial	Size	Color
		Precision ↑	Recall ↑	F1 ↑	Acc. ↑	Acc. ↑	Acc. ↑
×	×	78.81	59.44	66.58	30.74	26.75	18.78
×	✓	79.76	59.34	67.03	36.43	**30.34**	18.39
✓	×	**82.11**	**59.35**	**67.59**	**36.92**	**28.94**	**23.88**
✓	✓	**81.25**	**59.39**	**67.54**	**40.22**	27.74	**26.32**

Performance evaluation of LLMs

Model	Format ↑	Valid ↑	Correct ↑
Llama 1 [49]	67.5	46.0	38.5
Llama 2 [50]	98.5	84.0	63.5
GPT-3 [6]	**98.5**	97.5	83.5
GPT-4 [37]	**98.5**	**98.5**	**88.5**

Instructing text-to-image by chatGPT

Limitation

Limitation

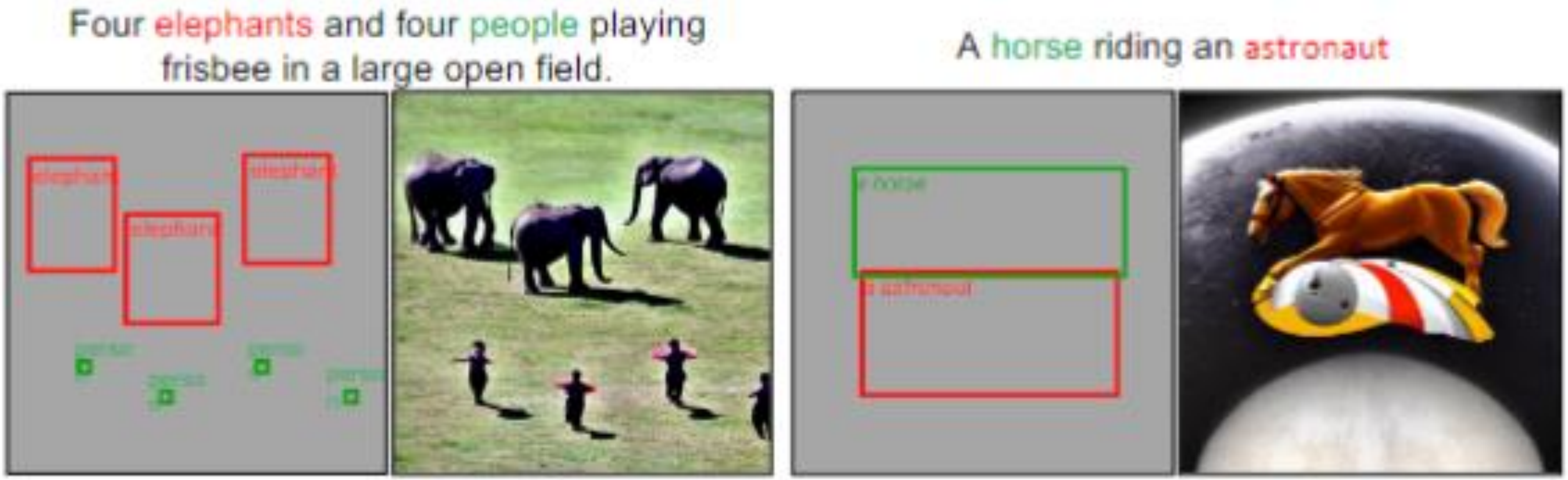


Figure 11. **The failures cases of our framework.** GPT-4 sometimes misinterprets object quantity or size and instances of the text-to-image model not aligning with GPT-4's layout

Thanks for Listening!