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

Biao Gong*✉ Siteng Huang* Yutong Feng Shiwei Zhang Yuyuan Li Yu Liu

Alibaba Group & Zhejiang University

CVPR SEATTLE, WA JUNE 17-21, 2024

Overview

Research Question

Can we calibrate the "layout hallucination" in pre-trained text-to-image generators in real-time without layout annotations?

Contributions

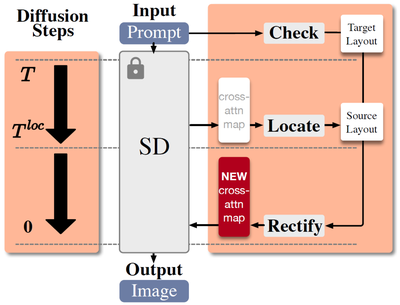
- devise a training-free layout calibration system *SimM* that intervenes in the generative process on the fly during inference time.
- present a benchmark *SimMBench* that compensates for the lack of superlative spatial relations in existing datasets.
- report both quantitative and qualitative results to demonstrate the effectiveness in automatically calibrating the layout inconsistencies.

Methodology

Preliminaries

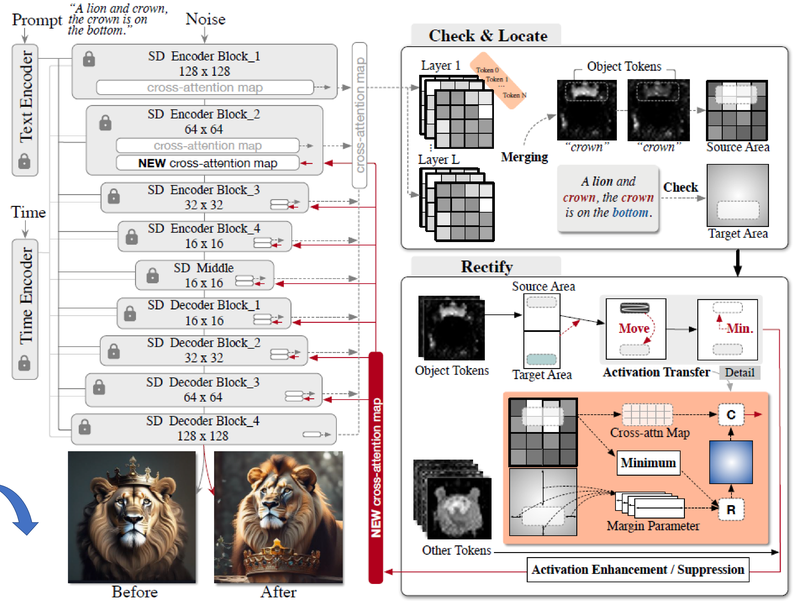
- Stable Diffusion leverages cross-attention layers to incorporate textual cues for the control of the image generation.
- For the object corresponding to the *k*-th token of the prompt, higher activations on the intermediate cross-attention maps indicate the approximate position where the object will appear.

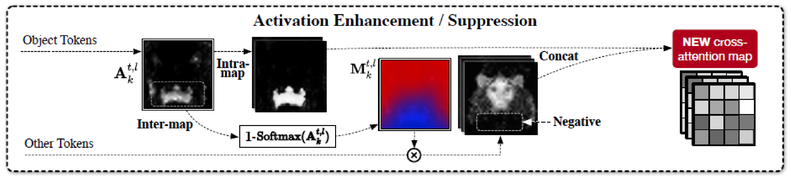
Our *SimM* system follows a "**check-locate-rectify**" pipeline:



Check

- detects the presence of object layout requirements in the textual prompts with predefined positional vocabulary.
- generates approximate target layout for each object by parsing the prompt and applying heuristic rules.
- assesses any discrepancies between the generated image and the specified layout requirements.





Locate

- identifies the source activated region for each object during the early denoising steps.

Rectify

- transfers the located activations to the target regions.
- adjusts the transferred activations with intra-/inter-map activation enhancement and suppression.

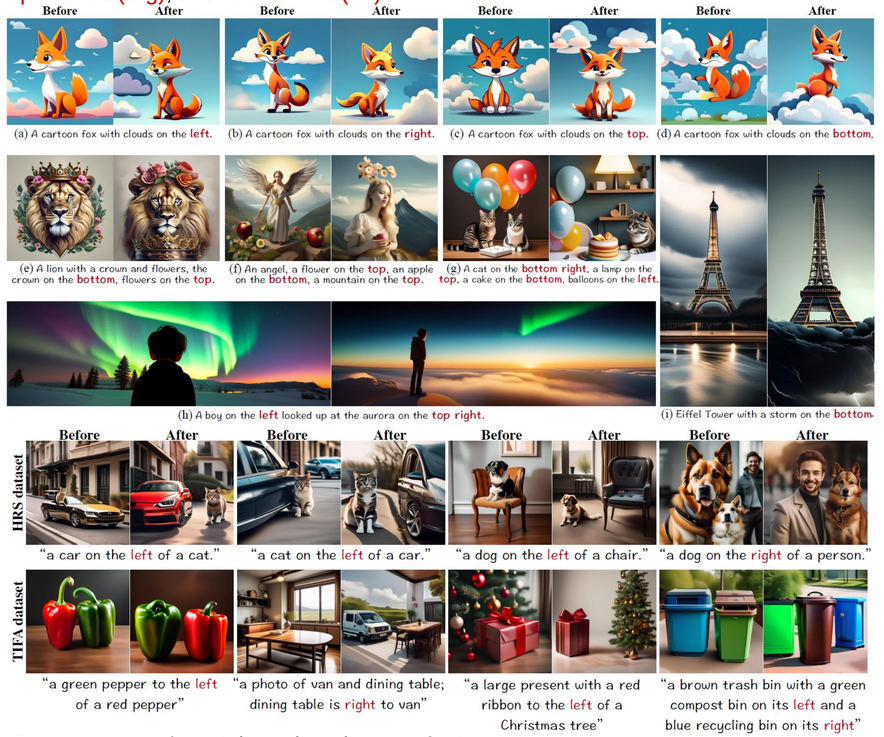
Main Results

SimMBench:

- 203 prompts, focusing on superlative relations
- 28 items, including single-word, phrase, and those with color

Methods	DrawBench [35] Accuracy	DrawBench [35] CLIP-Score	SimMBench Accuracy	SimMBench CLIP-Score
Stable Diffusion [32]	12.50	0.3267	4.25	0.3012
BoxDiff [40]	30.00	0.3239	24.08	**0.3032**
Layout-Guidance [6]	36.50	0.3354	25.50	0.3020
Attention-Refocusing [25]	43.50	0.3339	50.71	0.3017
SimM (Ours)	**53.00**	**0.3423**	**65.16**	0.3001

Calibration results across various position requirements (a-d), object quantities (e-g), and resolutions (h-i):



For more experimental results, please refer to our paper.