Control and Realism: Best of Both Worlds in Layout–to–Image without Training

Bonan Li[1,2], Yinhan Hu[1], Songhua Liu[2], Xinchao Wang[2]

[1]University of Chinese Academy of Science, [2]National University of Singapore

ICML International Conference On Machine Learning

University of Chinese Academy of Sciences

Introduction

Background

Recent advances in Text-to-Image (T2I) generation have profoundly revolutionized the vision landscape, facilitating the synthesis of highly authentic assets from textual prompts, e.g., text-driven Image-to-Image translation and video generation. Nevertheless, designing comprehensive prompts to meticulously control every aspect of an image can be both labor-intensive and time-consuming, posing challenges for efficient generation workflows.

Motivation

However, existing methods still suffer from:

- Imprecise localization;
- Unrealistic artifacts.

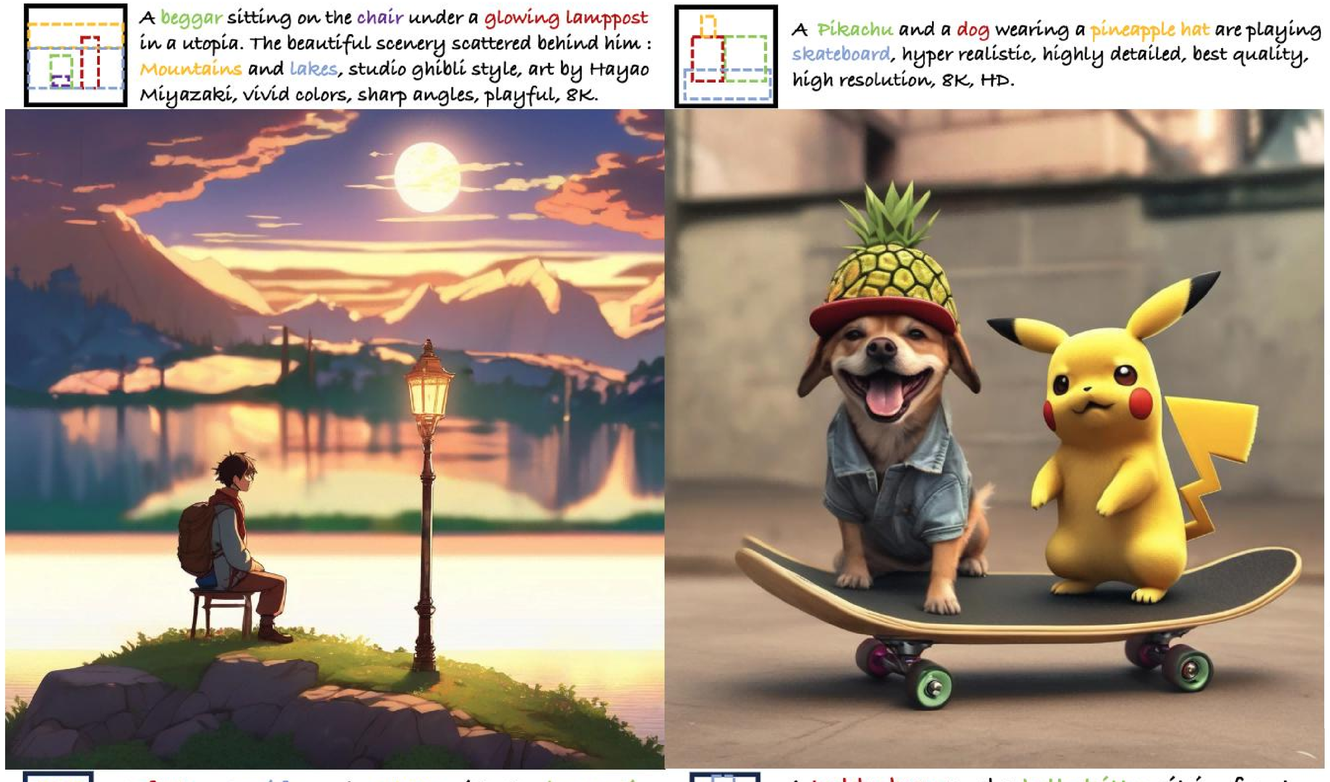
These issues primarily stem from:

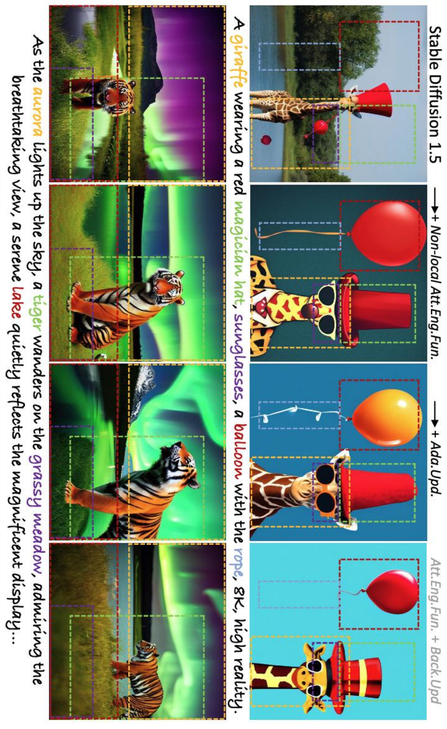
- Commonly used attention energy function introduces inherent spatial distribution biases, hindering objects from being uniformly aligned with layout instructions.
- Vanilla backpropagation update rule can cause deviations from the pre-trained domain, leading to out-of-distribution artifacts.

How to solve these issues:

- Non-local attention prior is explored to redistribute attention scores, facilitating objects to better conform to the specified spatial conditions.
- Langevin dynamicsbased adaptive update scheme as a remedy that promotes in-domain updating while respecting layout constraints.

Method





Model	COCO2014		
	AP_{50}↑	AP↑	CLIP-s↑
Att.Eng.Fun. + Back.Upd.	25.8	8.4	0.310
Non-local Att.Eng.Fun. + Back.Upd.	44.1	17.4	0.318
Att.Eng.Fun. + **Ada.Upd.**	36.7	14.9	0.324
Non-local Att.Eng.Fun. + Ada.Upd.	**49.2**	**19.7**	**0.327**