Progressive Compositionality in Text-to-Image Generative Models

Evans Xu Han, Linghao Jin, Xiaofeng Liu, Paul Pu Liang

Yale · USC · Massachusetts Institute of Technology

ICLR

MOTIVATION

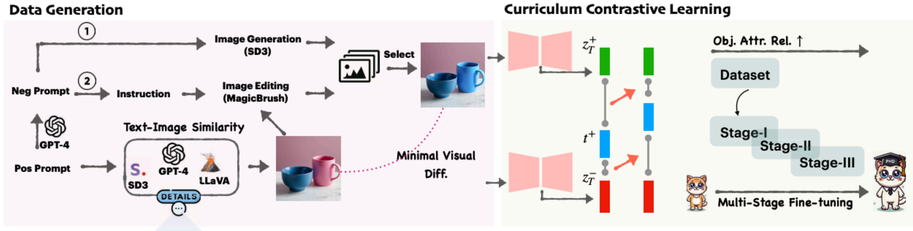
- Text-to-Image Models struggle to understand compositional relationships between objects and attributes, especially in complex settings.

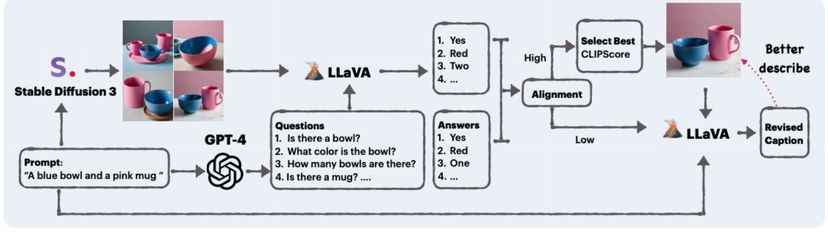
Introduction

- **Contrastive compositional dataset**. We introduce CONPAIR, a meticulously crafted compositional dataset consisting of high-quality contrastive images with minimal visual representation differences, covering a wide range of attribute categories.
- **EVOGEN: Curriculum contrastive learning.** We also incorporate curriculum contrastive learning into a diffusion model to improve compositional understanding.

METHODOLOGY

Step 1: Data Construction





Given a text prompt, we will generate one positive image and one negative image, by leveraging VQA model and LLMs to guarantee the faithfulness between generated image and text prompt.

Step 2: Curriculum Contrastive Fine-tuning

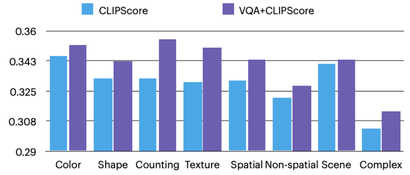
We apply contrastive loss during fine-tuning across three stages, from simple to complex. Given a positive text prompt, and a pair of contrastive images:

$$\mathcal{L} = -\log \frac{\exp(\text{sim}(h_t^+, f(t))/\tau)}{\exp(\text{sim}(h_t^+, f(t))/\tau) + \exp(\text{sim}(h_t^-, f(t))/\tau)}$$

Results

Text-Image Alignment in Dataset

Applying VQA checker consistently improves text-image alignment.



Alignment Evaluation

Contrastively Finetuning on synthesized dataset effectively improves alignment ability progressively across multi-stages.

Model	Attribute Binding			Object Relationship		Complex
	Color	Shape	Texture	Spatial	Non-Spatial	
STABLE V2 (Rombach et al., 2022)	50.65	42.21	49.22	13.42	30.96	33.86
CONPAIR	63.63	47.64	61.64	17.77	31.21	35.02
CONPAIR + Contra. Loss	69.45	54.39	67.72	20.21	32.09	38.14
CONPAIR + Contra. Loss + Multi-stage FT	**71.04**	**54.57**	**72.34**	**21.76**	**33.08**	**42.52**

User Study

We conducted a user study to complement our evaluation and provide a more intuitive assessment of EVOGEN's performance.