Progressive Compositionality in Text-to-Image Generative Models

Xu Han[1] **Linghao Jin**[2] **Xiaofeng Liu**[1] **Paul Pu Liang**[3]
[1]Yale University, [2]University of Southern California, [3]Massachusetts Institute of Technology
{xu.han.xh365, xiaofeng.liu}@yale.edu
linghaoj@usc.edu pplliang@mit.edu

Abstract

Despite the impressive text-to-image (T2I) synthesis capabilities of diffusion models, they often struggle to understand compositional relationships between objects and attributes, especially in complex settings. Existing solutions have tackled these challenges by optimizing the cross-attention mechanism or learning from the caption pairs with minimal semantic changes. However, *can we generate high-quality complex contrastive images that diffusion models can directly discriminate based on visual representations?* In this work, we leverage large-language models (LLMs) to compose realistic, complex scenarios and harness Visual-Question Answering (VQA) systems alongside diffusion models to automatically curate a contrastive dataset, ConPair, consisting of 15k pairs of high-quality contrastive images. These pairs feature minimal visual discrepancies and cover a wide range of attribute categories, especially complex and natural scenarios. To learn effectively from these error cases, i.e., hard negative images, we propose EvoGen, a new multi-stage curriculum for contrastive learning of diffusion models. Through extensive experiments across a wide range of compositional scenarios, we showcase the effectiveness of our proposed framework on compositional T2I benchmarks. The project page can be found at https://github.com/evansh666/EvoGen.

1 Introduction

The rapid advancement of text-to-image generative models (Saharia et al., 2022; Ramesh et al., 2022) has revolutionized the field of image synthesis, driving significant progress in various applications such as image editing (Brooks et al., 2023; Zhang et al., 2024), video generation (Brooks et al., 2024) and medical imaging (Han et al., 2024a). Despite their remarkable capabilities, state-of-the-art models such as Stable Diffusion (Rombach et al., 2022) and DALL-E 3 (Betker et al., 2023) still face challenges with *composing multiple objects into a coherent scene* (Huang et al., 2023; Liang et al., 2023; Majumdar et al., 2024). Common issues include *incorrect attribute binding, miscounting*, and *flawed object relationships* as shown in Figure 1. For example, when given the prompt "a red motorcycle and a yellow door", the model might incorrectly bind the colors to the objects, resulting in a yellow motorcycle.

Recent progress focuses on optimizing the attention mechanism within diffusion models to better capture the semantic information conveyed by input text prompts (Agarwal et al., 2023; Chefer et al., 2023; Pandey et al., 2023). For example, Meral et al. (2023) proposes manipulating the attention on objects and attributes as contrastive samples during test-time to optimize model performance. While more focused, the practical application of these methods still falls short of fully addressing attribute binding and object relationships. Other works advocate to develop compositional generative models to improve compositional performance as each constituent model captures distributions of an independent domain (Du & Kaelbling, 2024). However, such approach assumes a fixed prespecified structure to compose models, limiting generalization to new distributions.

In this paper, we argue that curriculum training is crucial to equip diffusion models with a fundamental understanding of compositionality. Given that existing models often struggle with even basic tasks (e.g., generating three cats when prompted with *"Two cats are playing"*) (Wang et al.,

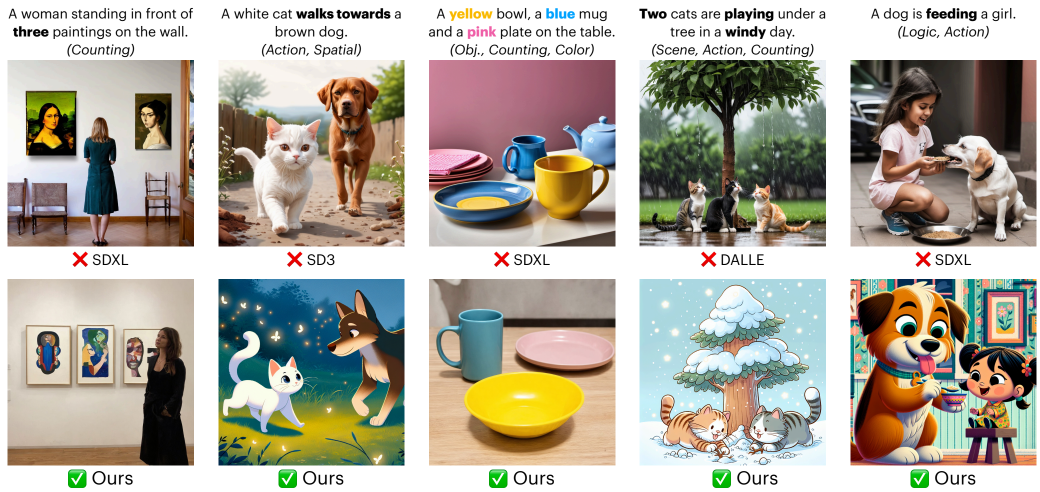

Figure 1: **Limited Compositionality Understanding in Diffusion Models**. Existing SOTA models such as SDXL, DALL-E 3 often fail to correctly compose objects and attributes. The bottom are images generated by our EvoGen.

2024), we progressively introduce more complex compositional scenarios during fine-tuning. This staged training strategy helps models build a solid foundation before tackling intricate cases before improving their performance on a wide range of compositional tasks.

Although many datasets exist for compositional generation (Wang et al., 2023; Feng et al., 2023), there remains a significant gap in datasets that offer a clear progression from simple to complex samples within natural and reasonable contexts. Moreover, creating high-quality contrastive image datasets is both costly and labor-intensive, especially given the current limitations of generative models in handling compositional tasks. To address this, we propose an automatic pipeline to generate faithful contrastive image pairs, which we find crucial for guiding models to focus on compositional discrepancies. In summary, our work can be summarized as follows:

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. By leveraging LLMs, we scale up the complexity of compositional prompts while maintaining a natural context design. Our dataset features faithful images generated by diffusion models, assisted by VQA systems to ensure accurate alignment with the text prompts.

EvoGen: Curriculum contrastive learning. We are the first work to incorporate curriculum contrastive learning into a diffusion model to improve compositional understanding. The process is broken into three streamlined sub-tasks: (1) learning single object-attribute composition, (2) mastering attribute binding between two objects, and (3) handling complex scenes with multiple objects. We conduct extensive experiments using the latest benchmarks and demonstrate that EvoGen significantly boosts the model's compositional understanding, outperforming most baseline generative methods.

2 Preliminary Background

2.1 Diffusion models

We implement our method on top of the state-of-the-art text-to-image (T2I) model, Stable Diffusion (SD) (Rombach et al., 2022). In this framework, an encoder $\mathcal{E}$ maps a given image $x \in \mathcal{X}$ into a spatial latent code $z = \mathcal{E}(x)$, while a decoder $\mathcal{D}$ reconstructs the original image, ensuring $\mathcal{D}(\mathcal{E}(x)) \approx x$.

A pre-trained denoising diffusion probabilistic model (DDPM) (Sohl-Dickstein et al., 2015; Ho et al., 2020) for noise estimation and a pre-trained CLIP text encoder (Radford et al., 2021) to

Dataset	# Samples	Contra. text	Contra. Image	Categories	Complex
DRAWBENCH (Saharia et al., 2022)	200	✗	✗	3 (color, spatial, action)	✓
CC-500 (Feng et al., 2023)	500	✗	✗	1 (color)	✗
ATTN-AND-EXCT (Chefer et al., 2023)	210	✗	✗	2 (color, animal obj.)	✗
T2I-COMPBENCH (Huang et al., 2023)	6000	✗	✗	6 (color, counting, texture, shape, (non-)spatial, complex)	✓
GEN-AI (Li et al., 2024a)	1600	✗	✗	8 (scene, attribute, relation, counting, comparison, differentiation, logic)	✓
ABC-6K (Feng et al., 2023)	6000	✓	✗	1 (color)	✗
WINOGROUNDT2I (Zhu et al., 2023)	22k	✓	✗	20 (action, spatial, direction, color, number, size, texture, shape, age, weight, manner, sentiment, procedure, speed, etc.)	✗
COMP. SPLITS (Park et al., 2021)	31k	✓	✓	2 (color, shape)	✗
WINOGROUND (Thrush et al., 2022)	400	✓	✓	5 (object, relation, symbolic, series, pragmatics)	✗
EQBEN (Wang et al., 2023)	250k	✓	✓	4 (attribute, location, object, count)	✗
ARO (Yuksekgonul et al., 2023)	50k	✓	✓	(relations, attributes)	✗
CONPAIR (ours)	15k	✓	✓	8 (color, counting, shape, texture, (non-)spatial relations, scene, complex)	✓

Table 1: The comparison of compositional T2I datasets. Contra. is the abbreviation of Contrastive. *Complex* refers the samples that have multiple objects and complicated attributes and relationships.

process text prompts into conditioning vectors $c(y)$. The DDPM model $\epsilon(\theta)$ is trained to minimize the difference between the added noise ϵ and the model's estimate at each timestep t,

$$\mathcal{L} = \mathbb{E}_{z\sim\mathcal{E}(x),y,\varepsilon\sim\mathcal{N}(0,1),t}\left[||\varepsilon - \varepsilon_\theta(z_t, t, c(y))||_2^2\right]. \tag{1}$$

During inference, a latent z_T is sampled from $\mathcal{N}(0, 1)$ and is iteratively denoised to produce a latent z_0. The denoised latent z_0 is then passed to the decoder to obtain the image $x' = \mathcal{D}(z_0)$.

2.2 COMPOSITIONAL DATASETS AND BENCHMARKS

The most commonly used data sets for object-attribute binding, including DRAWBENCH (Saharia et al., 2022), CC-500 (Feng et al., 2023) and ATTEND-AND-EXCITE (Chefer et al., 2023) construct text prompts by conjunctions of objects and a few of common attributes like *color* and *shape*. To more carefully examine how generative models work on each compositional category, recent work explores the disentanglement of different aspects of text-to-image compositionality. Huang et al. (2023) introduces T2I-COMPBENCH that constructing prompts by LLMs which covers six categories including *color, shape, textual, (non-)spatial relationships* and *complex compositions*; Recently, GEN-AI (Li et al., 2024a) collects prompts from professional designers which captures more enhanced reasoning aspects such as *differentiation, logic* and *comparison*.

Another line of work proposes contrastive textual benchmarks to evaluate the compositional capability of generative models. ABC-6K (Feng et al., 2023) contains contrast pairs by either swapping the order objects or attributes while they focus on negative text prompts with minimal changes. WINOGROUNDT2I (Zhu et al., 2023) contains 11K complex, high-quality contrastive sentence pairs spanning 20 categories. However, such benchmarks focus on text perturbations but do not have images, which have become realistic with the advancement of generative models.

Several benchmarks featuring contrastive image pairs have also been introduced. COMPOSITIONAL SPLITS C-CUB AND C-FLOWERS (Park et al., 2021) mainly focused on the color and shape attributes of birds and flowers, sourcing from Caltech-UCSD Birds (Wah et al., 2011), Oxford-102 (Flowers) (Nilsback & Zisserman, 2008). Thrush et al. (2022) curated WINOGROUND consists of 400 high-quality contrastive text-image examples. EQBEN (Wang et al., 2023) is an early effort to use Stable Diffusion to synthesize images to evaluate the equivariance of VLMs similarity, but it lacks more complex scenarios. Yuksekgonul et al. (2023) emphasizes the importance of hard negative samples and constructs negative text prompts in ARO by swapping different linguistic elements in the captions sourced from COCO and sampling negative images by the nearest-neighbor algorithm. However, it is not guaranteed the negative images found in the datasets truly match the semantic meaning of the prompts.

3 DATA CONSTRUCTION: CONPAIR

To address attribute binding and compositional generation, we propose a new high-quality contrastive dataset, CONPAIR. Next, we introduce our design principle for constructing CONPAIR. Each sample in CONPAIR consists of a pair of images (x^+, x^-) associated with a positive caption t^+.

Category	Stage-I	Stage-II
Shape	An **american football**. A **volleyball**.	An american football and a volleyball. A badminton ball and Frisbee.
Color	A**blue** backpack. A **red** backpack	A **blue** backpack and a **yellow** purse. A **yellow** purse and a **blue** backpack.
Counting	**Three** birds. **Two** birds.	Two **cats** and one **dog**. Two **dogs** and one **cat**.
Texture	A **plastic** toy. A **fluffy** toy.	A rubber **tire** and a glass **mirror**. A rubber **mirror** and a glass **tire**
Spatial	–	A **plate** on the right of a **bee**. A **bee** on the right of a **place**.
Non-spatial	A basketball player is **eating dinner**. A basketball player is **dancing**.	A **woman** is passing a tennis ball to a **man**. A **man** is passing a tennis ball to a **woman**.
Scene	A **snowy** night. A **rainy** night.	In a serene lake during a **thunderstorm**. In a serene lake on a **sunny day**.
Complex	**Two round** clock. **Three square** clock.	**Two fluffy dogs** are eating apples to the right of **a brown cat**. **A brown dog** are eating pears to the left of **two fluffy cats**.
	Stage-III	
Complex	Two **green** birds standing next to two **orange** birds on a **willow tree**. An **orange** bird standing next to three **green** birds on the **grass**.	
	A **man** wearing a **blue** hat is throwing an **american football** from the left to the right to a woman wearing a **green** pants on the playground during a **snowy day**. A **woman** wearing a **green** hat is throwing a **tennis ball** from the right to the left to a woman wearing a **blue** hat on the playground during a **rainy night**.	

Table 2: Examples of text prompts. Each sample has a positive (top) and a negative prompt (bottom).

3.1 Generating text prompts

Our text prompts cover eight categories of compositionality: *color, shape, texture, counting, spatial relationship, non-spatial relationship, scene*, and *complex*. To obtain prompts, we utilize the in-context learning capability of LLMs. We provide hand-crafted seed prompts as examples and predefined templates (e.g., *"A {color} {object} and a {color} {object}."*) and then ask GPT-4 to generate similar textual prompts. We include additional instructions that specify the prompt length, no repetition, etc. In total, we generate 15400 positive text prompts. More information on the text prompt generation is provided in the appendix A.

To generate a negative text prompt t^-, we use GPT-4 to perturb the specified attributes or relationships of the objects for Stage-I data. In Stage-II, we either swap the objects or the attributes, depending on which option makes more sense in the given context. For complex sentences, we prompt GPT-4 to construct contrastive samples by altering the attributes or relationships within the sentences. Table 2 presents our example contrastive text prompts.

3.2 Generating contrastive images

Minimal Visual Differences. Our key idea is to generate contrastive images that are minimally different in visual representations. By "minimal," we mean that, aside from the altered attribute/relation, other elements in the images remain consistent or similar. In practice, we source negative image samples in two ways: 1) generate negative images by prompting negative prompts to diffusion models; 2) edit the positive image by providing instructions (e.g., change motorcycle color to red) using MagicBrush (Zhang et al., 2024), as shown at the left of Figure 2.

Text-Image Alignment. The high-level objective of ConPair is to generate positive images that faithfully adhere to the positive text guidance, while the corresponding negative images do not align with the positive text, despite having minimal visual differences from the positive images. As the quality of images generated by diffusion-based T2I generative models varies significantly (Karthik et al., 2023), we first generate 10-20 candidate images per prompt. However, how to select the most faithful image is difficult. Existing automatic metrics like CLIPScore are not always effective at comparing the faithfulness of images when they are visually similar. To address this, we propose

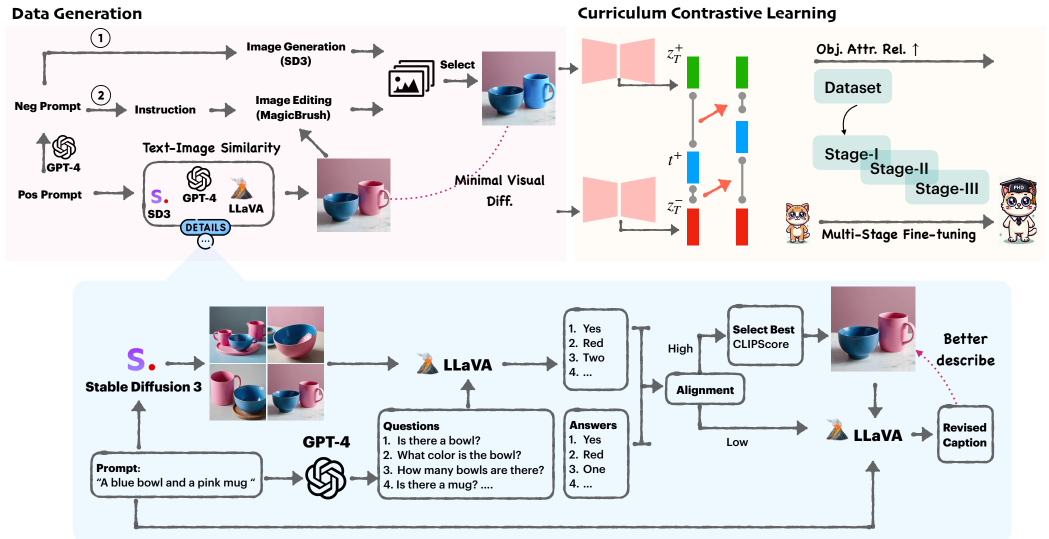

Figure 2: **EvoGen Framework**. **Data generation pipeline** (left) **and curricumlum contrastive learning** (right). **Quality control of image geneartion** (bottom): Given a prompt, SD3 generates multiple candidate images, which are evaluated by LLaVA. We select the best image by alignment and CLIPScore. If the alignment score is low, we prompt LLaVA to describe the image as a new revised caption based on the generated image.

decomposing each text prompt into a set of questions using an LLM and leverage the capabilities of VQA models to rank candidate images by their alignment score, as illustrated in Figure 2 (bottom) [1]. Note the correct answers can be directly extracted from the prompts. Intuitively, we consider an image a success if all the answers are correct or if the alignment is greater than θ_{align} for certain categories, such as *Complex*. After getting aligned images, we select the best image by automatic metric (e.g., CLIPScore).

Empirically, we find this procedure fails to generate faithful images particularly when the prompts become *complex*, as limited by the compositionality understanding of existing generative models, which aligns with the observations of Sun et al. (2023). In response to such cases–i.e., the alignment scores for all candidate images are low–we introduce an innovative reverse-alignment strategy. Instead of simply discarding low-alignment images, we leverage a VLM to dynamically revise the text prompts based on the content of the generated images. By doing so, we generate new captions that correct the previous inaccuracies while preserving the original descriptions, thereby improving the alignment between the text and image.

Image-Image Similarity. Given each positive sample, we generate 20 negative images and select the one with the highest similarity to the corresponding positive image, ensuring that the changes between the positive and negative image pairs are minimal. In case of *color* and *texture*, we use image editing rather than generation, as it delivers better performance for these attributes. Han et al. (2024b) proposes that human feedback plays a vital role in enhancing model performance. For quality assurance, 3 annotators randomly manually reviewed the pairs in the dataset and filtered 647 pairs that were obviously invalid.

4 EvoGen: Curriculum Contrastive Fine-tuning

A common challenge in training models with data of mixed difficulty is that it can overwhelm the model and lead to suboptimal learning (Bengio et al., 2009). Therefore, we divide the dataset into three stages and introduce a simple but effective multi-stage fine-tuning paradigm, allowing the model to gradually progress from simpler compositional tasks to more complex ones.

[1] Examples of decomposed questions are provided in the Appendix A.3

Figure 3: **Contrastive dataset** examples. Each pair includes a positive image generated from the given prompt (left) and a negative image that is semantically inconsistent with the prompt (right), differing only minimally from the positive image.

Stage-I: Single object. In the first stage, the samples consist of a single object with either a specific attribute (e.g., shape, color, quantity, or texture), a specific action, or within a simple static scene. The differences between the corresponding negative and positive images are designed to be clear and noticeable. For instance, "*A man is **walking***" vs. "*A man is **eating***", where the actions differ significantly, allowing the model to easily learn to distinguish between them.

Stage-II: Object compositions. We compose two objects with specified interactions and spatial relationships. An example of *non-spatial relationship* is "*A **woman** chases a **dog***" vs. "*A yellow **dog** chases a **woman***." This setup helps the models learn to differentiate the relationships between two objects.

Stage-III: Complex compositions. To further complicate the scenarios, we propose prompts with complex compositions of attributes, objects, and scenes. Data in this stage can be: 1) contain more than two objects; 2) assign more than two attributes to each object, or 3) involve intricate relationships between objects.

Ultimately, our goal is to equip the model with the capability to inherently tackle challenges in compositional generation. Next, we discuss how to design the contrastive loss during fine-tuning at each stage. Given a positive text prompt t, a generated positive image x^+, and corresponding negative image x^-, the framework comprises the following three major components:

Diffusion Model. The autoencoder converts the positive image and negative image to latent space as z_0^+ and z_0^-. The noisy latent at timestep t is represented as z_t^+ and z_t^-. The encoder of the noise estimator ϵ_θ is used to extract feature maps z_{et}^+ and z_{et}^- respectively.

Model	Attribute Binding			Object Relationship		Complex
	Color	Shape	Texture	Spatial	Non-Spatial	
SD v1.4 (Rombach et al., 2022)	37.65	35.76	41.56	12.46	30.79	30.80
DALL-E 2 (Ramesh et al., 2022)	57.00	55.00	63.74	13.00	30.00	37.00
SDXL (Podell et al., 2023)	64.00	54.00	36.45	20.00	31.00	41.00
Composable v2 (Liu et al., 2023)	40.63	32.99	36.45	8.00	29.80	28.98
Structured v2 (Feng et al., 2023)	49.90	42.18	49.00	13.86	31.11	33.55
Attn-Exct v2 Chefer et al. (2023)	64.00	45.17	59.63	14.55	31.09	34.01
GORs (Huang et al., 2023)	66.03	47.85	62.87	18.15	31.93	33.28
PixArt-α (Chen et al., 2023)	68.86	55.82	70.44	20.82	31.79	41.17
MARS (He et al., 2024)	69.13	54.31	71.23	19.24	32.10	40.49
DALL-E 3 (Betker et al., 2023)	77.85	62.05	70.36	28.65	30.03	37.73
SD v2.1 (Rombach et al., 2022)	50.65	42.21	49.22	13.42	30.96	33.86
EvoGen-SD v2.1 (Ours)	71.04	54.57	72.34	21.76	33.08	42.52
SD3-Medium (Esser et al., 2024)	80.63	58.48	75.24	34.88	31.17	47.72
EvoGen- SD3-Medium (Ours)	**83.08**	**62.83**	**77.59**	**36.95**	**33.12**	**49.07**

Table 3: Alignment evaluation on T2I-CompBench. The best results are in **bold**

Method	Basic						Advanced					
	Attribute	Scene	Relation			Avg	Count	Differ	Compare	Logical		Avg
			Spatial	Action	Part					Negate	Universal	
SD v2.1	0.75	0.77	0.72	0.72	0.69	0.74	0.66	0.63	0.61	0.50	0.57	0.58
SD-XL Turbo	0.79	0.82	0.77	0.78	0.76	0.79	0.69	0.65	0.64	**0.51**	0.57	0.60
DeepFloyd-IF	0.82	0.83	0.80	0.81	0.80	0.81	0.69	0.66	0.65	0.48	0.57	0.60
SD-XL	0.82	0.84	0.80	0.81	0.81	0.82	0.71	0.67	0.64	0.49	0.57	0.60
Midjourney v6	0.85	0.88	0.86	0.86	0.85	0.85	0.75	0.73	0.70	0.49	0.64	0.65
SD3-Medium	0.86	0.86	0.87	0.86	0.88	0.86	0.74	0.77	0.72	0.50	**0.73**	0.68
DALL-E 3	**0.91**	**0.91**	0.89	0.89	**0.91**	**0.90**	0.78	0.76	0.70	0.46	0.65	0.65
EvoGen- SD3-Medium (Ours)	0.89	0.88	**0.90**	**0.91**	0.88	0.89	**0.80**	**0.79**	**0.73**	**0.51**	**0.73**	**0.72**

Table 4: Gen-AI Benchmark Results.

Projection head. We apply a small neural network projection head $g(\cdot)$ that maps image representations to the space where contrastive loss is applied. We use a MLP with one hidden layer to obtain $h_t = g(z_{et}) = W^{(2)}\sigma(W^{(1)}(z_{et}))$.

Contrastive loss. For the contrastive objective, we utilize a variant of the InfoNCE loss (van den Oord et al., 2019), which is widely used in contrastive learning frameworks. This loss function is designed to maximize the similarity between the positive image and its corresponding text prompt, while minimizing the similarity between the negative image and the same text prompt. The loss for a positive-negative image pair is expressed as follows:

$$\mathcal{L} = -\log \frac{\exp(\text{sim}(h_t^+, f(t))/\tau)}{\exp(\text{sim}(h_t^+, f(t))/\tau) + \exp(\text{sim}(h_t^-, f(t))/\tau)} \tag{2}$$

where τ is a temperature parameter, $f(\cdot)$ is CLIP text encoder, sim function represents cosine similarity:

$$\text{sim}(u, v) = \frac{u^T \cdot v}{\|u\|\|v\|} \tag{3}$$

This encourages the model to distinguish between positive and negative image-text pairs.

5 Experiments and Discussions

5.1 Implementation Details

Experimental Setup In an attempt to evaluate the faithfulness of generated images, we use GPT-4 to decompose a text prompt into a pair of questions and answers, which serve as the input of our VQA model, LLaVA v1.5 (Liu et al., 2024). Following previous work (Huang et al., 2023;

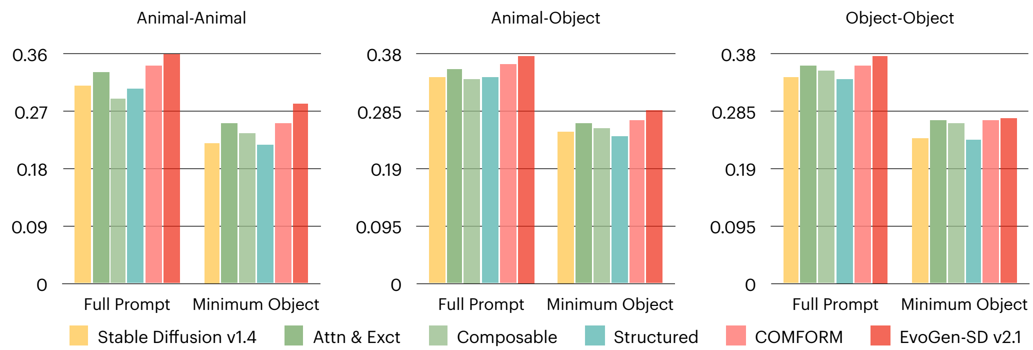

Figure 4: Average CLIP image-text similarities between the text prompts and the images generated by different models. The *Full Prompt* Similarity considers full text prompt. *Minimum Object* represents the minimum of the similarities between the generated image and each of the two object prompts. Example of this benchmark is in Appendix C.

Feng et al., 2023), we finetune and evaluate EVOGEN on SD v2.1 (Rombach et al., 2022) and SD3-Medium (Esser et al., 2024).

Baselines We compare our results with several state-ofthe-art methods, including trending open-sourced T2I models that trained on large training data, SD v1.4, SD v2.1 and SD3-Medium, DALL-E 2 (Ramesh et al., 2022) and SDXL (Podell et al., 2023). ComposableDiffusion v2 (Liu et al., 2023) is designed for conjunction and negation of concepts for pretrained diffusion models. StructureDiffusion v2 (Feng et al., 2023), Divide-Bind (Li et al., 2024b) and Attn-Exct v2 (Chefer et al., 2023) are designed for attribute binding for pretrained diffusion models. GORs (Huang et al., 2023) finetunes SD v2.1 with selected samples and rewards. PixArt-α (Chen et al., 2023) incorporates cross-attention modules into the Diffusion Transformer. MARS (He et al., 2024) adapts from auto-regressive pre-trained LLMs for T2I generation tasks.

Evaluation Metrics To quantitatively assess the efficacy of our approach, we comprehensively evaluate our method via three primary metrics: 1) compositionality on T2I-CompBench (Huang et al., 2023), 2) GenAI-Bench Li et al. (2024a) and 3) color-object compositionality prompts (Chefer et al., 2023).

5.2 Performance Comparison and Analysis

Alignment Assessment. To examine the quality of CONPAIR, we measure the alignment of the positive image and texts using CLIP similarity. Figure 5 compares directly selecting the best image based on CLIPScore with our pipeline, which leverages a VQA model to guide image generation. These results confirm that our approach consistently improves image faithfulness across all categories with VQA assistance during image generation and demonstrate CONPAIR contains high-quality image-text pairs.

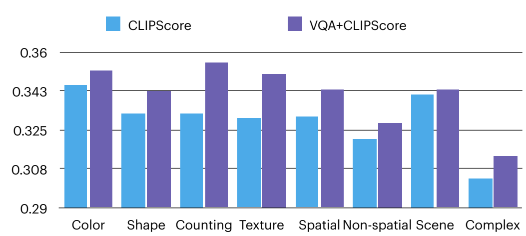

Figure 5: Average CLIP similarity of image-text pairs in CONPAIR.

Benchmark Results Beyond the above evaluation, we also assess the alignment between the generated images using EVOGEN and text condition on T2I-Compbench. As shown in Table 3, EVOGEN achieves SOTA performance across all categories. In particular, we observe significant improvements in the *Non-spatial* and *Complex* categories. We hypothesize this is primarily attributed to Stage-III training, where high-quality contrastive samples with complicated compositional components are leveraged to achieve superior alignment capabilities. We further evaluate EVOGEN

Model	Attribute Binding			Object Relationship		Complex
	Color	Shape	Texture	Spatial	Non-Spatial	
SD v2.1 (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 (EVOGEN-SD v2.1)	**71.04**	**54.57**	**72.34**	**21.76**	**33.08**	**42.52**

Table 5: Ablation on T2I-CompBench. CONPAIR refers to directly finetune SD v2.1 on CONPAIR.

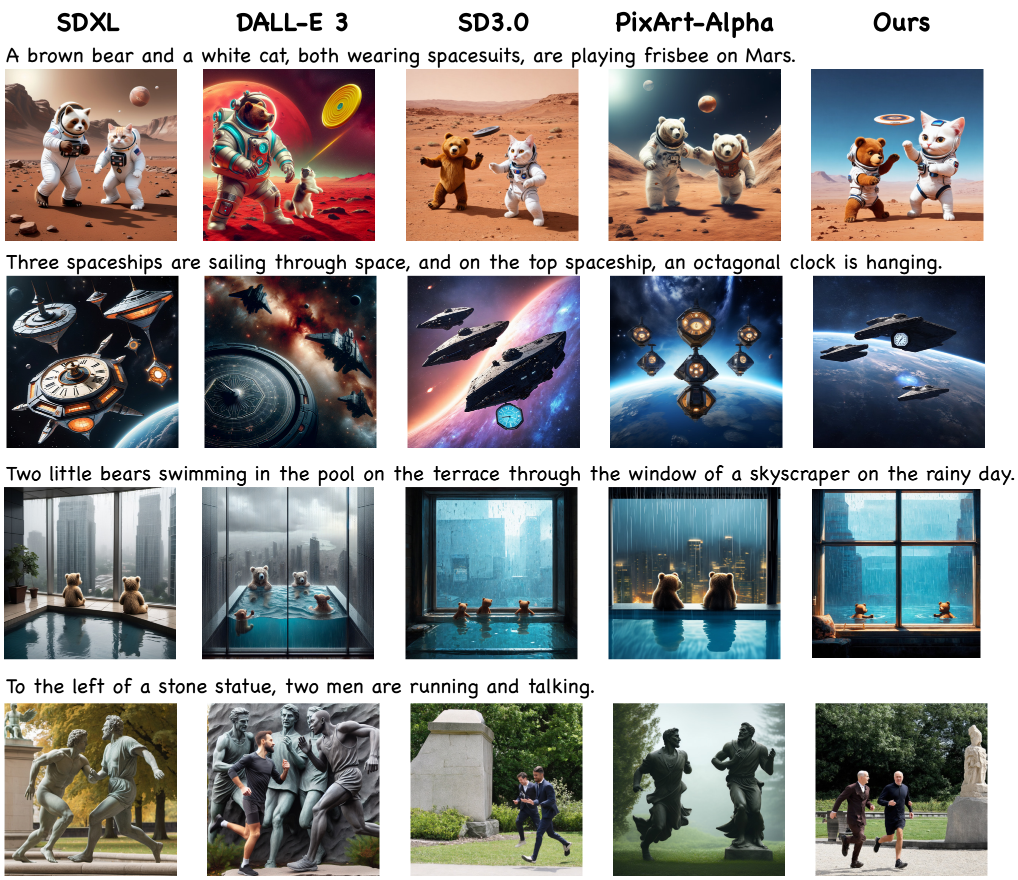

Figure 6: Qualitative comparison between EVOGEN and other SOTA T2I modes with different prompts.

on the Gen-AI benchmark. As indicated in Table 4, EVOGEN performs best on all the *Advanced* prompts, although it exhibits relatively weaker performance in some of the basic categories compared to DALL-E 3.

Figure 4 presents the average image-text similarity on the benchmark proposed by Chefer et al. (2023), which evaluates the composition of objects, animals, and color attributes. Compared to other diffusion-based models, our method consistently outperforms in both *full* and *minimum* similarities across three categories, except for the minimum similarity on Object-Object prompts. These results demonstrate the effectiveness of our approach.

Ablation Study We conduct ablation studies on T2I-CompBench by exploring three key design choices. First, we assess the effectiveness of our constructed dataset, CONPAIR, by fine-tuning SD v2.1 directly using CONPAIR. As shown in Table 5, our results consistently outperform the baseline evaluation on SD v2.1 across all categories, demonstrating that our data generation pipeline is effective. Next, we validate the impact of our contrastive loss by comparing it with fine-tuning without this loss. The contrastive loss improves performance in the attribute binding category, though it has less impact on object relationships and complex scenes. We hypothesize this is because attribute

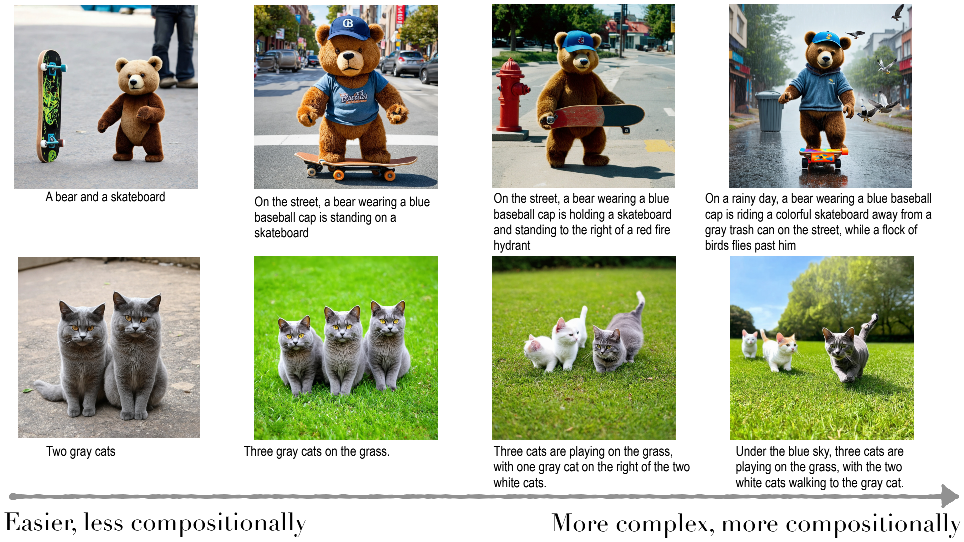

Figure 7: Examples of EvoGen for complex compositionality.

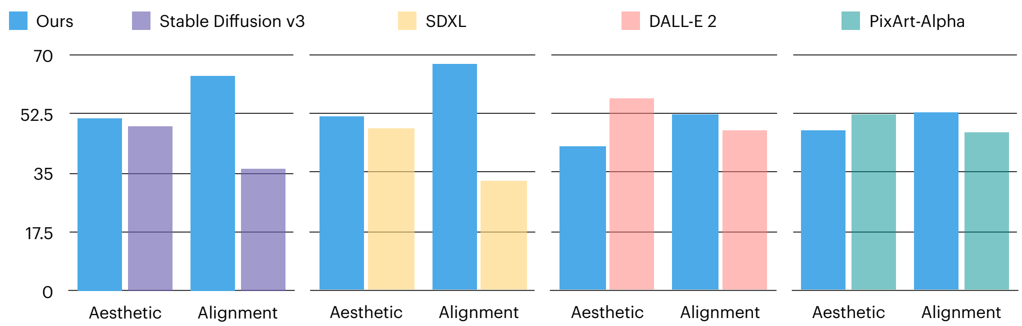

Figure 8: User study on 100 randomly selected prompts from Feng et al. (2023). The ratio values indicate the percentages of participants preferring the corresponding model.

discrepancies are easier for the model to detect, while relationship differences are more complex. Finally, applying the multi-stage fine-tuning strategy leads to further improvements, particularly in the *Complex* category, suggesting that building a foundational understanding of simpler cases better equips the model to handle more intricate scenarios.

Qualitative Evaluation Figure 6 presents a side-by-side comparison between EvoGen and other state-of-the-art T2I models, including SDXL, DALL-E 3, SD v3 and PixArt-α. EvoGen consistently outperforms the other models in generating accurate images based on the given prompts. SDXL frequently generates incorrect actions and binds attributes to the wrong objects. DALL-E 3 fails to correctly count objects in two examples and misses attributes in the first case. SD v3 struggles with counting and attribute binding but performs well in generating actions. PixArt-α is unable to handle attributes, spatial relationships, and fails to count objects accurately in the second prompt.

Next, we evaluate how our approach handles complex compositionality, as shown in Figure 7. Using the same object, "bear" and "cat," we gradually increase the complexity by introducing variations in attributes, counting, scene settings, interactions between objects, and spatial relationships. The generated results indicate that our model effectively mitigates the attribute binding issues present in existing models, demonstrating a significant improvement in maintaining accurate compositional relationships.

User Study We conducted a user study to complement our evaluation and provide a more intuitive assessment of EvoGen's performance. Due to the time-intensive nature of user studies involving human evaluators, we selected top-performing comparable models—DALLE-2, SD v3, SDXL, and

PixArt-α—all accessible through APIs and capable of generating images. As shown in Figure 8, the results demonstrate EvoGen's superior performance in alignment, though the aesthetic quality may be slightly lower compared to other models.

6 Conclusion

In this work, we present EvoGen, a curriculum contrastive framework to overcome the limitations of diffusion models in compositional text-to-image generation, such as incorrect attribute binding and object relationships. By leveraging a curated dataset of positive-negative image pairs and a multi-stage fine-tuning process, EvoGen progressively improves model performance, particularly in complex scenarios. Our experiments demonstrate the effectiveness of this method, paving the way for more robust and accurate generative models.

7 Limitation

Despite the effectiveness of our current approach, there are a few limitations that can be addressed in future work. First, our dataset, while comprehensive, could be further expanded to cover an even broader range of compositional scenarios and object-attribute relationships. This would enhance the model's generalization capabilities. Additionally, although we employ a VQA-guided image generation process, there is still room for improvement in ensuring the faithfulness of the generated images to their corresponding prompts, particularly in more complex settings. Refining this process and incorporating more advanced techniques could further boost the alignment between the text and image.

8 Reproducibility

We have made efforts to ensure that our method is reproducible. Appendix A provides a description of how we construct our dataset. Especially, Appndix A.1 and A.2 presents how we prompt GPT-4 and use predefined template to generate text prompts of our dataset. Appendix A.3 provides an example how we utilize VQA system to decompose a prompt into a set of questions, and answers. Appendix B provides the details of implementation, to make sure the fine-tuning is reproducible.

References

Aishwarya Agarwal, Srikrishna Karanam, K. J. Joseph, Apoorv Saxena, Koustava Goswami, and Balaji Vasan Srinivasan. A-star: Test-time attention segregation and retention for text-to-image synthesis. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2283–2293, 2023. URL https://api.semanticscholar.org/CorpusID:259252450.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. ICML '09, pp. 41–48, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585161. doi: 10.1145/1553374.1553380. URL https://doi.org/10.1145/1553374.1553380.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions, 2023. URL https://arxiv.org/abs/2211.09800.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. URL https://openai.com/research/video-generation-models-as-world-simulators.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2301.13826.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023. URL https://arxiv.org/abs/2310.00426.

Yilun Du and Leslie Kaelbling. Compositional generative modeling: A single model is not all you need, 2024. URL https://arxiv.org/abs/2402.01103.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. URL https://arxiv.org/abs/2403.03206.

Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis, 2023. URL https://arxiv.org/abs/2212.05032.

Xu Han, Fangfang Fan, Jingzhao Rong, and Xiaofeng Liu. Fair text to medical image diffusion model with subgroup distribution aligned tuning, 2024a. URL https://arxiv.org/abs/2406.14847.

Xu Han, Felix Yu, Joao Sedoc, and Benjamin Van Durme. Baby bear: Seeking a just right rating scale for scalar annotations, 2024b. URL https://arxiv.org/abs/2408.09765.

Wanggui He, Siming Fu, Mushui Liu, Xierui Wang, Wenyi Xiao, Fangxun Shu, Yi Wang, Lei Zhang, Zhelun Yu, Haoyuan Li, Ziwei Huang, LeiLei Gan, and Hao Jiang. Mars: Mixture of auto-regressive models for fine-grained text-to-image synthesis, 2024. URL https://arxiv.org/abs/2407.07614.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation, 2023. URL https://arxiv.org/abs/2307.06350.

Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. If at first you don't succeed, try, try again: Faithful diffusion-based text-to-image generation by selection, 2023. URL https://arxiv.org/abs/2305.13308.

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. Genai-bench: Evaluating and improving compositional text-to-visual generation, 2024a. URL https://arxiv.org/abs/2406.13743.

Yumeng Li, Margret Keuper, Dan Zhang, and Anna Khoreva. Divide & bind your attention for improved generative semantic nursing, 2024b. URL https://arxiv.org/abs/2307.10864.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations and trends in multimodal machine learning: Principles, challenges, and open questions, 2023. URL https://arxiv.org/abs/2209.03430.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2024. URL https://arxiv.org/abs/2310.03744.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. Compositional visual generation with composable diffusion models, 2023. URL https://arxiv.org/abs/2206.01714.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, Karmesh Yadav, Qiyang Li, Ben Newman, Mohit Sharma, Vincent-Pierre Berges, Shiqi Zhang, Pulkit Agrawal, Yonatan Bisk, Dhruv Batra, Mrinal Kalakrishnan, Franziska Meier, Chris Paxton, Alexander Sax, and Aravind Rajeswaran. Openeqa: Embodied question answering in the era of foundation models. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16488–16498, 2024. URL https://api.semanticscholar.org/CorpusID:268066655.

Tuna Han Salih Meral, Enis Simsar, Federico Tombari, and Pinar Yanardag. Conform: Contrast is all you need for high-fidelity text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2312.06059.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pp. 722–729, 2008. URL https://api.semanticscholar.org/CorpusID:15193013.

Rohan Pandey, Rulin Shao, Paul Pu Liang, Ruslan Salakhutdinov, and Louis-Philippe Morency. Cross-modal attention congruence regularization for vision-language relation alignment, 2023. URL https://arxiv.org/abs/2212.10549.

Dong Huk Park, Samaneh Azadi, Xihui Liu, Trevor Darrell, and Anna Rohrbach. Benchmark for compositional text-to-image synthesis. In *NeurIPS Datasets and Benchmarks*, 2021. URL https://api.semanticscholar.org/CorpusID:244906179.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023. URL https://arxiv.org/abs/2307.01952.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. URL https://arxiv.org/abs/2204.06125.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

Jiao Sun, Deqing Fu, Yushi Hu, Su Wang, Royi Rassin, Da-Cheng Juan, Dana Alon, Charles Herrmann, Sjoerd van Steenkiste, Ranjay Krishna, and Cyrus Rashtchian. Dreamsync: Aligning text-to-image generation with image understanding feedback, 2023. URL https://arxiv.org/abs/2311.17946.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality, 2022. URL https://arxiv.org/abs/2204.03162.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. URL https://arxiv.org/abs/1807.03748.

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Tan Wang, Kevin Lin, Linjie Li, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Equivariant similarity for vision-language foundation models, 2023. URL https://arxiv.org/abs/2303.14465.

Zhenyu Wang, Enze Xie, Aoxue Li, Zhongdao Wang, Xihui Liu, and Zhenguo Li. Divide and conquer: Language models can plan and self-correct for compositional text-to-image generation, 2024. URL https://arxiv.org/abs/2401.15688.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it?, 2023. URL https://arxiv.org/abs/2210.01936.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing, 2024. URL https://arxiv.org/abs/2306.10012.

Xiangru Zhu, Penglei Sun, Chengyu Wang, Jingping Liu, Zhixu Li, Yanghua Xiao, and Jun Huang. A contrastive compositional benchmark for text-to-image synthesis: A study with unified text-to-image fidelity metrics, 2023. URL https://arxiv.org/abs/2312.02338.

A ConPair Data Construction

A.1 Text prompts generation

Here, we design the template and rules to generate text prompts by GPT-4 as follows:

- *Color*: Current state-of-the-art text-to-image models often confuse the colors of objects when there are multiple objects. Color prompts in Stage-I follow fixed sentence template *"A {color} {object}."* and *"A {color} {object} and a {color} {object}."* for Stage-II.
- *Texture*: Following Huang et al. (2023), we emphasize in the GPT-4 instructions to require valid combinations of an object and a textural attribute. The texture prompts follows the template *"A {texture} {object}."* for Stage-I and *"A {texture} {object} and a {texture} {object}."* for Stage-II.
- *Shape*: We first generate objects with common geometric shapes using fixed template *"A {shape} {object}."* for Stage-I and *"A {shape} {object} and a {shape} {object}."* for Stage-II. Moreover, we ask GPT-4 to generate objects in the same category but with different shapes, e.g., American football vs. Volleyball, as contrastive samples.
- *Counting*: Counting prompts in Stage-I follows fixed sentence template *"{count} {object}."* and *"{count} {object} and {count} {object}."* for Stage-II.
- *Spatial Relationship*: Given predefined spatial relationship such as *next to, on the left, etc*, we prompt GPT-4 to generate a sentence in a fixed template as *"{object} {spatial} {object}."* for Stage-II.
- *Non-spatial Relationship*: Non-spatial relationships usually describe the interactions between two objects. We prompt GPT-4 to generate text prompts with non-spatial relationships (e.g., actions) and arbitrary nouns. We guarantee there is only one object in the sentence for Stage-I, and two objects in Stage-II. We also find generative models fails to understand texts like *"A woman is passing a ball to a man"*. It's hard for the model to correctly generate the directions of actions. We specially design prompts like this.
- *Scene*: We ask GPT-4 to generate scenes such as weather, place and background. For Stage-I, the scene is simple, less than 5 words (e.g., on a rainy night.); For Stage-II, scenes combine weather and background or location (e.g., in a serene lake during a thunderstorm.).
- *Complex:* Here, we refer to prompts that either contain more than two objects or assign more than two attributes to each object, or involve intricate relationships between objects. We first manually curate 10 such complex prompts, each involving multiple objects bound to various attributes. These manually generated prompts serve as a context for GPT-4 to generate additional natural prompts that emphasize compositionality. The complex cases in Stage-II will be two objects with more attributes; Stage-III involves more objects.

A.2 Negative Text Prompts Generation

We apply in-context learning, everytim we prompt GPT-4 to generate negative cases, we give 5-10 example test prompts each time, and make sure the generation is not repetitive, within certain lengths.

- In Stage-I, we prompt GPT-4 to change the attribute of the object such as color, shape, texture, counting, action, or scene, with instruction the differences should be noticeable.
- In Stage-II, we either swap the objects or attributes and let GPT-4 to validate the swapped text prompts. For complex cases, we generate negative text by asking GPT-4 to change the attributes/relationship/scenes.
- In Stage-III, we carefully curate complicated examples with 3-6 objects, each object have 1-3 attributes, with negative prompts change attributes, actions and spatial relationships, scenes. And we prompt GPT-4 with such examples.

A.3 VQA Assistance

Instruction for QA Generation. Given an image description, generate one or two multiple-choice questions that verify if the image description is correct. Table 6 shows an example of a generated prompt and QA.

Prompt	Question	Answer
A brown bear and a white cat, both wearing spacesuits, are playing frisbee on Mars	Is there a bear?	Yes
	Is there a cat?	Yes
	What color is the bear?	Brown
	What color is the cat?	White
	Does the bear wear a spacesuit?	Yes
	Does the cat wear a spacesuit?	Yes
	Is the bear playing the frisbee?	Yes
	Is the cat playing the frisbee?	Yes
	Where are they playing?	Mars

Table 6: VQA generated questions from a prompt.

Next, we illustrate how we prompt VQA to revise the caption when alignment scores of all candidate images are low. Given a generated image and a original text prompt, we prompt LLaVA *"Given the original text prompt describing the image, identify any parts that inaccurately reflect the image. Then, generate a revised text prompt with correct descriptions, making minimal semantic changes."*. At the same time, we will give example of revised caption pairs to VQA for in-context learning.

A.4 Data Statistics

The dataset is organized into three stages, each progressively increasing in complexity. In Stage-I, the dataset includes simpler tasks such as Shape (500 samples), Color (800), Counting (800), Texture (800), Non-spatial relationships (800), and Scene (800), totaling 4,500 samples. Stage-II introduces more complex compositions, with each category—including Shape, Color, Counting, Texture, Spatial relationships, Non-spatial relationships, and Scene—containing 1,000 samples, for a total of 7,500 samples. Stage-III represents the most complex scenarios, with fewer but more intricate samples. We also include some simple cases like Stage-I and II, each contain 200 samples, while the Complex category includes 2,000 samples, totaling 3,400 samples. Across all stages, the dataset contains 15,400 samples, providing a wide range of compositional tasks for model training and evaluation. Figure 9 show more examples of images in our dataset.

	Stage-I	Stage-II	Stage-III	Total
Shape	500	1000	200	1700
Color	800	1000	200	2000
Counting	800	1000	200	2000
Texture	800	1000	200	2000
Spatial	-	1000	200	1200
Non-spatial	800	1000	200	2000
Scene	800	1000	200	2000
Complex	-	500	2000	2500

Table 7: Corpus Statistics.

B Training Implementation Details

We implement our apprpoach upon stable Diffuion v2.1 and we employ the pre-trained text encoder from the CLIP ViT-L/14 model. The VAE encoder is frozen during training. The resolution is 768, the batch size is 16, and the learning rate is 3e-5 with linear decay.

A yellow dog running with a woman

A square clock hanging on the wall

A green apple and a yellow pear

Four hobbits are preparing to cross a hexagonal time tunnel in an underground volcanic cave

A golden vase and a clear glass

Cupid is playing with a pink rabbit on white clouds with a cotton candy texture

A cat is **chasing** a woman

A pyramid **in the desert** under the sky

A man is walking on the street

A garden with various flowers, five garden gnomes—three wearing red hats and two wearing green hats—a birdbath in the middle, and a wooden bench on the left, with two butterflies flying above

Three wolves in the foggy weather and a bright moon in the sky.

A green car is parked between two blue motorcycles, and a person wearing a red jacket is standing to the rear left of the car, holding a briefcase.

Figure 9: Contrastive Images in CONPAIR

C ATTN & EXCT BENCHMARK PROMPT EXAMPLES

The benchmark protocol we follow comprises structured prompts ‘a [animalA] and a [animalB]’, ‘a [animal] and a [color][object]’, ‘a [colorA][objectA] and a [colorB][objectB]’ . Table 8 demonstrate the results of average CLIP similarities between text prompts and captions generated by BLIP for Stable Diffusion-based methods on this benchmark. EVOGEN outperform those models on three categories.

Model	Animal-Animal	Animal-Obj	Obj-Obj
STABLE v1.4 (Rombach et al., 2022)	0.76	0.78	0.77
COMPOSABLE v2 (Liu et al., 2023)	0.69	0.77	0.76
STRUCTURED v2 (Feng et al., 2023)	0.76	0.78	0.76
ATTN-EXCT v2 (Chefer et al., 2023)	0.80	0.83	0.81
CONFORM (Meral et al., 2023)	0.82	0.85	0.82
Ours	0.84	0.86	0.85

Table 8: Attn-Exct benchmark Results.

D QUALITATIVE RESULTS

Figure 10 presents more comparison between EVOGEN and other state-of-the-art T2I models, including SDXL, DALL-E 3, SD v3 and PixArt-α.

Figure 10: Qualitative Results.