Separate-and-Enhance: Compositional Finetuning for Text2Image Diffusion Models

Zhipeng Bao
CMU
USA
zbao@cs.cmu.edu

Yijun Li
Adobe Research
USA
yijli@adobe.com

Krishna Kumar Singh
Adobe Research
USA
krishsin@adobe.com

Yu-Xiong Wang
UIUC
USA
yxw@illinois.edu

Martial Hebert
CMU
USA
hebert@cs.cmu.edu

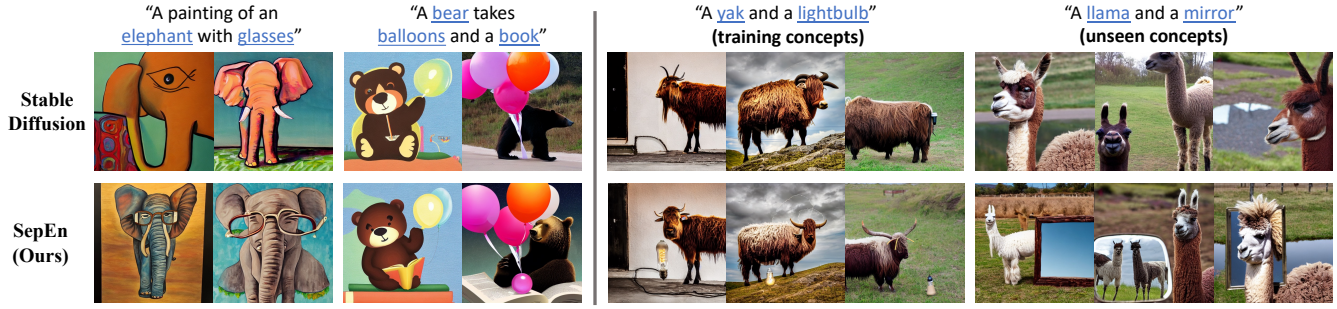

Figure 1: Visual comparisons between Stable Diffusion [Rombach et al. 2022] and our method. Different from previous *inference-based* methods, we propose a compositional finetuning algorithm for Text2Image diffusion models that can improve the text-image alignment and scale up to a large collection of concepts. Left: compositional finetuning for individual concepts. We are able to generate higher-quality images that are more aligned with the text input. Right: joint compositional finetuning with a large collection of concepts. After finetuning, the model keeps a high compositional capacity even for unseen novel concepts.

ABSTRACT

Despite recent significant strides achieved by diffusion-based Text-to-Image (T2I) models, current systems are still less capable of ensuring decent compositional generation aligned with text prompts, particularly for the multi-object generation. In this work, we first show the fundamental reasons for such misalignment by identifying issues related to low attention activation and mask overlaps. Then we propose a compositional finetuning framework with two novel objectives, the Separate loss and the Enhance loss, that reduce object mask overlaps and maximize attention scores, respectively. Unlike conventional test-time adaptation methods, our model, once finetuned on critical parameters, is able to directly perform inference given an arbitrary multi-object prompt, which enhances the scalability and generalizability. Through comprehensive evaluations, our model demonstrates superior performance in image realism, text-image alignment, and adaptability, significantly surpassing established baselines. Furthermore, we show that training our model with a diverse range of concepts enables it to generalize effectively to novel concepts, exhibiting enhanced performance compared to models trained on individual concept pairs.

CCS CONCEPTS

• **Computing methodologies** → **Image processing**.

KEYWORDS

Image Generation, Diffusion Models

ACM Reference Format:
Zhipeng Bao, Yijun Li, Krishna Kumar Singh, Yu-Xiong Wang, and Martial Hebert. 2024. Separate-and-Enhance: Compositional Finetuning for Text2Image Diffusion Models. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers '24 (SIGGRAPH Conference Papers '24), July 27–August 01, 2024, Denver, CO, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3641519.3657527

SIGGRAPH Conference Papers '24, July 27–August 01, 2024, Denver, CO, USA

ACM ISBN 979-8-4007-0525-0/24/07
https://doi.org/10.1145/3641519.3657527

1 INTRODUCTION

Human cognition possesses a remarkable capacity for compositional understanding, allowing us to focus on, differentiate between, and even conceptualize novel objects [Tomasello 2009]. Mirroring this capability within machine vision systems, especially in the realm of generative modeling such as Text-to-Image (T2I) synthesis [Ramesh et al. 2021; Rombach et al. 2022] which seeks to produce

photo-realistic images that coherently represent given textual descriptions, has gained significant attention in the computer vision and graphics communities. However, even cutting-edge diffusion-based [Ho et al. 2020] T2I models, including the notable Stable Diffusion [Rombach et al. 2022], struggle with challenges when representing multiple objects with varying attributes, such as different shapes, sizes, and colors as shown at the top of Figure 2.

To understand the underlying reasons, we visualize the text-image correspondence for two examples with cross-attention masks at the bottom of Figure 2. Through our analysis, we find two primary factors for the observed text-image misalignment: (1) the attention activation scores for certain objects are notably low and (2) the attention masks corresponding to diverse objects exhibit substantial overlap. Potential causes for these phenomena include the dominance of certain object classes among others, and the rarity of certain combinations during the model training.

Previous research has identified similar reasons, and endeavored to address them [Agarwal et al. 2023; Chefer et al. 2023; Huang et al. 2023; Li et al. 2023]. Nonetheless, these approaches primarily focus on a single facet of the problem: either amplifying attention activation [Chefer et al. 2023; Li et al. 2023] or reducing attention overlaps [Agarwal et al. 2023; Huang et al. 2023]. Our empirical observations indicate that solutions focusing solely on one aspect yield limited enhancements to the compositionality of T2I models, as supported by our results in Table 1 and Figure 8. Therefore, to better address these challenges, we propose two novel objectives respectively: the *Separate loss* designed to mitigate the Intersection of Union (IoU) of multiple objects and prevent their merging into a singular entity; and the *Enhance loss*, which seeks to maximize the attention activation scores associated with each object.

In light of the proposed objectives above, one question emerges: **How to best execute compositional finetuning?** A considerable portion of previous work [Agarwal et al. 2023; Chefer et al. 2023; Li et al. 2023] employs a strategy akin to test-time adaptation. Specifically, they retain the weights of pretrained T2I models and refine only the latent features for each pair of new concepts. Notable shortcomings of this type of approach include: (1) it fails to truly improve the compositional capacity of diffusion models as it remains training-free; (2) it results in longer inference time owing to the adaptation performed during testing; and most critically, (3) it is unable to scale up to *multiple* concepts, which lacks the **generalizability** for novel unseen concepts. To address these limitations, we directly apply our objectives to finetune the diffusion model, aiming at enhancing the inherent compositional ability of pretrained T2I models. Furthermore, a deep analysis of each function within the text-image attention modules enables us to selectively finetune a specific subset of parameters, the Key mapping functions (Section 3). This strategy leads to an overall lightweight finetuning process and makes it possible for our method to *generalize to larger scales*.

To validate our mechanism, we conduct a comprehensive experimental evaluation. Firstly, we evaluate our model with several individual prompt pairs. Our model achieves a much higher success rate of text-image alignment and more realistic image generation compared with the baselines including the Stable Diffusion model [Saharia et al. 2022] and its following work [Chefer et al. 2023; Liu et al. 2022] (see Figure 1, left). Subsequently, large-scale experiments on a collection of 220 concepts from ImageNet-21K [Deng et al. 2009] indicate that our method manages to simultaneously process multiple concept pairs, even performing better than variants training on each single pair. Notably, our resulting model from large-scale finetuning shows a great generalization capacity for unseen concepts (see Figure 1, right). Thirdly, we additionally ablate the attributions of each of the two objectives to the final promising results. Lastly, we show that our finetuned model not only improves compositional generation of multiple objects, but also retains comparable performance for single-object synthesis.

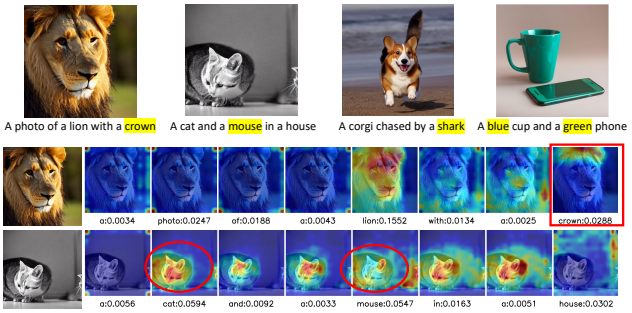

Figure 2: (Top) Failure cases and (Bottom) causes of Stable Diffusion [Rombach et al. 2022]. Even state-of-the-art T2I models fail when representing multiple objects with varying attributes. Two primary factors, demonstrated by the bottom examples respectively, include: (1) low attention activation for certain objects and (2) the attention masks overlap.

In conclusion, our contributions are threefold: **(1)** We analyze the underlying factors responsible for the compositional misalignment in T2I diffusion models. Subsequent to this analysis, we introduce two novel objectives: the *Separate loss* to decouple the object masks and the *Enhance loss* to maximize the attention activation scores of each object; **(2)** Instead of performing *test-time adaptation*, by carefully designing a **compositional finetuning** scheme, our model can greatly improve the composition capacity of T2I diffusion models for both individual concept pairs and a large collection of concepts, outperforming established baselines; **(3)** Crucially, by learning jointly from an extensive collection of concepts, our model learns a *global representation* of multiple concepts which better models the relationships between multiple concepts. Therefore it not only outperforms variants that are finetuned for individual concept pairs, but also demonstrates remarkable **generalization capabilities** for novel concepts. Our project page is available at https://github.com/zpbao/SepEn.

2 RELATED WORK

Text-to-Image (T2I) Synthesis. T2I synthesis aims to generate realistic and semantically consistent images from textual descriptions [Reed et al. 2016]. A lot of previous efforts have made significant advancements in this space based on generative adversarial networks (GANs) [Bau et al. 2021; Goodfellow et al. 2014; Xu et al. 2018; Zhang et al. 2017; Zhu et al. 2019], variational autoencoder (VAEs) [Ramesh et al. 2021], and diffusion models [Gu et al. 2022; Nichol et al. 2021; Ramesh et al. 2022; Rombach et al. 2022; Saharia et al. 2022; Yu et al. 2022]. Among them, diffusion models have

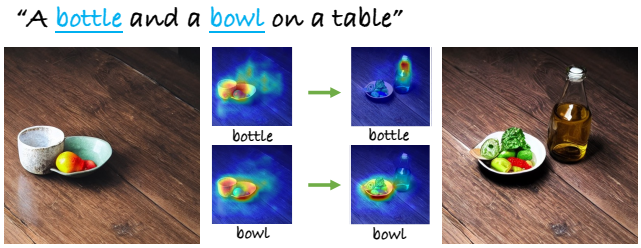

Figure 3: Example of before and after applying our separate loss $\mathcal{L}_{\text{Sep}}$. Left: attention maps from SD. The *bottle* attention falls onto the *bowl*. Right: attention maps from our model where the attention of *bottle* and *bowl* are better separated so that each concept is validly generated.

Figure 4: Example of before and after applying our enhance loss $\mathcal{L}_{\text{En}}$. Left: attention maps from SD. The attention score of *mouse* is lower than that of *cat* on the bottom right region, resulting in generating a cat-like mouse. Right: attention maps from our model where the attention activation score of *mouse* is enhanced so that it can be correctly generated.

emerged as a promising approach and show state-of-the-art generation quality. These models are built upon principles of heat diffusion and anisotropic diffusion to process and generate images [Perona and Malik 1990; Weickert 1998]. Combining diffusion and neural networks results in powerful generative models such as the Denoising Diffusion Probabilistic Model (DDPM) [Ho et al. 2020] and the Score-Based Generative Model (SBGM) [Song et al. 2021]. The text information is usually introduced via the cross-attention [Vaswani et al. 2017] modules during the denoising process. These methods showcase the adaptability and effectiveness of diffusion models in various text-based synthesis and editing applications.

Compositional Synthesis with Diffusion Models. Existing T2I models have been observed to be less capable of generating multiple objects described in prompts, which essentially indicates their poor compositional ability and thus causes text-image misalignment. Therefore, two main branches of methods are developed to enhance the compositional text-image alignment for T2I models. The first branch of work tackles the task in a style of test-time adaptation by running attention guidance [Epstein et al. 2023; Hertz et al. 2022; Hong et al. 2023; Rassin et al. 2023; Wang et al. 2024], tweaking attention masks [Agarwal et al. 2023; Chefer et al. 2023; Li et al. 2023], or refining latent representations [Brooks et al. 2023; Liu et al. 2022]. Among them, Attend-and-Excite [Chefer et al. 2023] encourages the attention activation for all objects to be as strong as possible. A more recent work, A-Star [Agarwal et al. 2023], designs novel loss functions to split the objects from one another. Though these two methods share a similar spirit as our approach, one major limitation is that they handle concepts pair by pair which lacks the **generalizability** to novel concepts. Another work, SynGen [Rassin et al. 2023], also focuses on improving the object-attribute matching for T2I diffusion models during inference time through additional linguistic guidance. All these methods are test-time-adaptation-based solutions that increase runtime and require parameter tuning per pair. The other type of methods further finetune the pretrained diffusion model with additional structured input or supervision, such as masks [Huang et al. 2023; Wang et al. 2022; Zhang and Agrawala 2023], bounding boxes [Ma et al. 2023; Zheng et al. 2023], selected images [Kim et al. 2023; Kumari et al. 2023], and external vision-language models [Feng et al. 2023; Singh and Zheng 2023]. While more signals beyond text would help reduce the text-image misalignment issue, our work focuses on improving the compositional ability of the T2I model itself given the text prompt only. We *do not* require any additional supervision during both training and inference time. Our goal is that our finetuned model not only improves the compositional generation of multiple objects, but also maintains comparable performance for single-object synthesis, so that eventually only one single model is needed for direct inference from just the text prompt input.

3 PROPOSED METHOD

3.1 Preliminary

Stable Diffusion. We build our model upon the state-of-the-art Stable Diffusion (SD) [Rombach et al. 2022]. Different from the basic pixel-based diffusion models [Ho et al. 2020; Song et al. 2020], SD operates in the latent space of the autoencoder rather than the image space. First, an encoder $\mathcal{E}$ is trained to map a given image $x \in \mathcal{X}$ into a spatial latent code $z = \mathcal{E}(x)$. A decoder $\mathcal{D}$ is then tasked with reconstructing the input image such that $\mathcal{D}(\mathcal{E}(x)) \approx x$.

SD first pretrains the large-scale autoencoder, and then it further trains a DDPM model that operates over the learned latent space to produce a denoised version of a noise input z_t at timestep t. During the denoising process, the diffusion model can be conditioned on an additional input vector such as text, layout, or semantic maps [Rombach et al. 2022]. Concretely, in SD, the conditional input is a prompt embedding produced by a pretrained CLIP text encoder [Radford et al. 2021]. Given conditional embedding $c(y)$ on the prompt y, the DDPM model ϵ_θ is trained to minimize the loss,

$$\mathcal{L} = \mathbb{E}_{z\sim\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}||\epsilon - \epsilon_\theta(z_t, t, c(y))||^2. \quad (1)$$

In summary, at each timestep t, ϵ_θ is tasked with correctly removing the noise ϵ added to the latent code z, given the noised latent z_t, timestep t, and conditioning encoding $c(y)$. ϵ_θ is a network with UNet architecture [Ronneberger et al. 2015] consisting of self-attention and cross-attention layers, as discussed below.

Text Condition. SD uses text embedding to guide image synthesis through cross-attention modules, which contain a self-attention layer followed by a cross-attention layer in each group. Concretely, the latent vector input of the i^{th} layer of the UNet, z_t^i, works as the query of the cross-attention, and the projections of the text embedding $c(y) \in \mathbb{R}^{N\times d_{\text{text}}}$, where N is the length of the text sequence and d_{text} is the embedding dimension, are used as the key

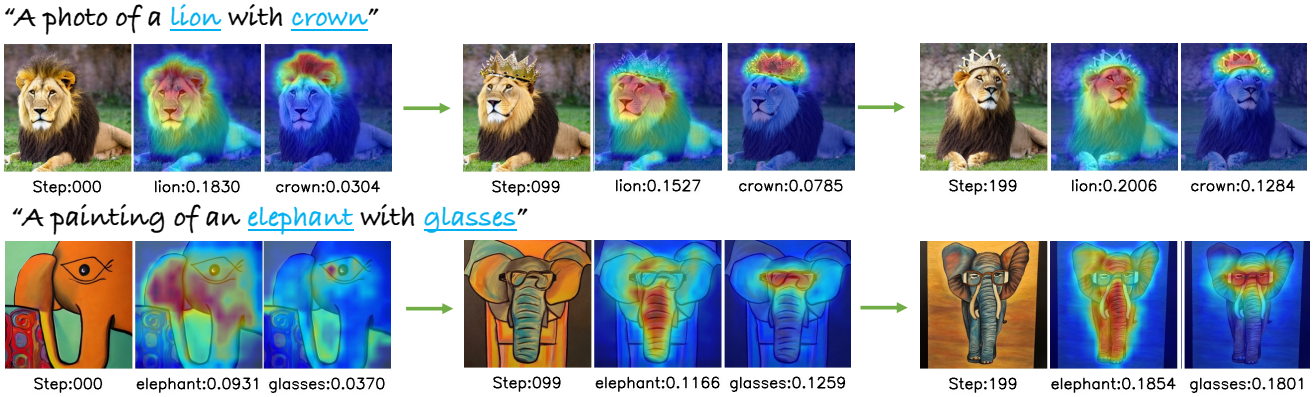

Figure 5: Example of change of generated images and object attention during different finetuning steps. The attention of two concepts in each row are gradually split and amplified until forming the final aligned image with high quality, showing the effectiveness of our proposed SepEn algorithm.

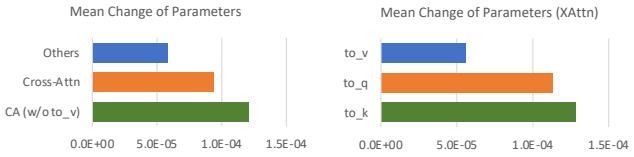

Figure 6: Average parameter changes for the whole network (left) and inside cross-attention modules (right) during finetuning. The cross-attention modules are more sensitive to finetuning, especially for the key mapping functions.

and value of the multi-head attention (MHA) [Vaswani et al. 2017] module: $z_t^{i'} = \text{MHA}(q(z_t^i),\ k(c(y)),\ v(c(y)))$.

By running the attention operation, we can also get the attention masks $M_t \in \mathbb{R}^{N\times(p\times p)}$ for further processing, where $p \in \{64, 32, 16, 8\}$ is the spatial resolution of the latent feature. For our method, we operate on attention masks under the resolution of 16×16. For all the following sections, we note $M_t \in \mathbb{R}^{N\times(16\times16)}$ as the averaged attention masks for all the cross attention modules with the resolution of 16×16.

3.2 Separate-and-Enhance Pipeline

As illustrated in Figure 2, two essential problems for the compositional misalignment are attention mask overlaps and low activation scores. We propose the Separate loss and the Enhance loss to tackle these two problems respectively as follows.

Separate Loss. As discussed in Section 1, one reason for the compositional misalignment is the overlapping attention masks. As further shown in Figure 3 (left), the original Stable Diffusion model fails to generate the *bottle* in the given prompt. By visualizing the attention masks, we find that the primary reason is that the attention of the *bottle* falls onto the "bowl" region in the image.

Therefore, we propose the Separate loss which aims at splitting object masks from binding to the same object region with the hope that salient objects should have as few overlaps as possible (refer to Figure 3, right). Concretely, during training, we randomly sample a timestep t and obtain the attention masks for all the K objects $\{M_t^i\}_{i=1}^K$. The Separate loss minimizes the maximum interaction of union for all the pixels as follows:

$$\mathcal{L}_{\text{Sep}} = max(\frac{\prod_{i=1}^K M_t^i}{\sum_{i=1}^K M_t^i}), \tag{2}$$

where $\prod$ indicates pixel-level product.

Enhance Loss. Another reason for the text-image misalignment is the low activation for certain objects. From Figure 4 (left), we see that the attention score of *mouse* is lower than that of *cat* on the bottom right region, resulting in generating a cat-like mouse. To solve this problem, we propose the Enhance loss aiming to highlight the attention activation score for all the objects with the hope that the synthesized images should have saliency regions for all the objects from the input prompts (refer to Figure 4, right).

Before computing the loss, inspired by Chefer et al. [2023], we first filter the attention masks with a Gaussian smooth kernel [Babaud et al. 1986] to obtain the smoothed version of the masks $\{\tilde{M}_t^i\}_{i=1}^K$. Next, the Enhance loss amplifies the attention of the lowest-scored concept by minimizing

$$\mathcal{L}_{\text{En}} = 1 - min(max(\tilde{M}_t^1), \cdots, max(\tilde{M}_t^K)). \tag{3}$$

3.3 Optimization

Tuned parameters. To efficiently and effectively perform the compositional finetuning, inspired by Custom Diffusion [Kumari et al. 2023], we conducted a pilot experiment by finetuning the whole UNet architecture with 20 pairs of concepts. The average parameter changes for different modules are shown in Figure 6. We find that the parameters in the cross-attention module are more sensitive to finetuning (left of Figure 6) compared with the other modules. Therefore, we decided to only finetune the parameters in the cross-attention modules. Moreover, we discovered that not all the parameters in the cross-attention modules should be tuned. We first show the general pipeline of the attention module with query (latent feature z and mapping function Q), key (prompt e_t and mapping function K), and value (e_t and mapping function V) as follows:

$$M = Q(z)K(e_t);\ z_{\text{out}} = \text{softmax}(M)V(e_t). \tag{4}$$

It is noted from Equation 4 that the Q function projects the input noisy latent, which has few connections to the misalignment issue;

Figure 7: Qualitative comparisons of our model and baselines. The images synthesized by our method have the best compositional alignment compared with the other baselines. Meanwhile, our method also maintains good visual quality for the generated images compared with the A & E baseline [Chefer et al. 2023].

Table 1: Results for the individual prompt evaluation. Gray values indicate publicly reported numbers. Our method has the best compositionality regarding Average Similarity Score and Success Rate. Meanwhile, we also have a lower FID than A & E, showing the promising quality of our generated images.

Method	FID (↓)	Average Similarity Score (↑)	Success Rate (↑)
StableDiffusion	32.96	0.742±0.091	0.209
+ Composable [Liu et al. 2022]	-	0.71	-
+ Structure [Feng et al. 2023]	-	0.77	-
+ A & E [Chefer et al. 2023]	45.65	0.793±0.088	0.383
+ A-Star [Agarwal et al. 2023]	-	0.83	-
+ SepEn	**36.85**	0.809 ± 0.086	**0.410**
+ SepEn (TTA)	41.74	**0.834 ± 0.081**	**0.441**

and the V function represents the feature embedding learned by the stable diffusion model to form the final latent vector for the VAE decoder, which we also want to maintain. Therefore, we select only the key mapping function K as the target parameter group to finetune, leading to a lightweight overall finetuning strategy. This conclusion is also consistent with the results of Figure 6 (right) where the key mapping function is the most sensitive one inside the cross-attention modules while the value mapping functions are the least. We also showcase in Section 4.3 that only finetuning the K function is the optimal choice compared with other variants.

Normalization term. As the proposed loss functions may lead to a distribution shift of the pretrained SD model, especially for large-scale finetuning, we also add a standard normalization term similar to Equation 1:

$$\mathcal{L}_{\text{norm}} = \sum_i \mathbb{E}_{z^i \sim \mathcal{E}(x^i), y^i, \epsilon \sim \mathcal{N}(0,1), t} ||\epsilon - \epsilon_\theta(z_t^i, t, c(y^i))||^2, \quad (5)$$

where (x^i, y^i) is the image-prompt pair sampled with frozen pretrained stable diffusion. The final loss objective is:

$$\mathcal{L}_{\text{final}} = \lambda_{\text{n}} \mathcal{L}_{\text{norm}} + \lambda_{\text{D}} \mathcal{L}_{\text{En}} + \lambda_{\text{E}} \mathcal{L}_{\text{Sep}}, \quad (6)$$

where λ_{n}, λ_{E}, and λ_{D} are weight factors.

Synergy of the two objectives. Notice that the two objectives tackle different aspects of text-image misalignment while they are not totally disentangled: splitting objects from overlapping with each other helps to better enhance the low-activated ones, while ensuring all the objects have high activation also helps to better detect the overlap in reverse. The synergy of the two objectives jointly improves the compositional capacity of T2I models. In Figure 5, we show the change of images and object attention during different finetuning steps. The attention of both *elephant* and *glasses* are gradually split and amplified until forming the final aligned image.

4 EXPERIMENTAL RESULTS

Baselines. We compare our model with three other diffusion-based, state-of-the-art T2I models. **SD** [Rombach et al. 2022] is a powerful T2I model trained on large-scale LAION-5B [Schuhmann et al. 2022] with billions of annotated pairs. The other baselines and our method all build upon this pretrained model. **Attend-and-Excite (A & E)** [Chefer et al. 2023] aims to maximize the attention value of the target nouns during the inference time, while **A-Star** [Agarwal et al. 2023] mitigates the attention mask overlaps. We report the original scores of A-Star due to a lack of open-source implementation. We also additionally report the results of Composable Diffusion [Liu et al. 2022] and Structure Diffusion [Feng et al. 2023] on single-prompt evaluations for reference.

Evaluation prompts. We design two groups of evaluation protocols: *individual prompts evaluation* and *large-scale evaluation*. For the individual prompts evaluation, we use the same group of test prompts from Chefer et al. [2023] that contains 276 prompts in three types: animal-animal, animal-object, and object-object. For the large-scale evaluation, we follow the process of Chefer et al. [2023] and select 220 concepts (110 animals and 110 objects) from ImageNet-21 [Deng et al. 2009] categories. We randomly select 10 animals and 10 objects as held-out categories. All of the compared methods do not require additional annotated data for finetuning, thereby no datasets are used.

Metrics. For the quantitative measurement, we adopt three metrics: (1) The **FID score** measures the realism of the generated images, by computing the Fréchet distance between two Gaussians fitted to feature representations of the source images and the target images [Dowson and Landau 1982; Parmar et al. 2022]. Given that

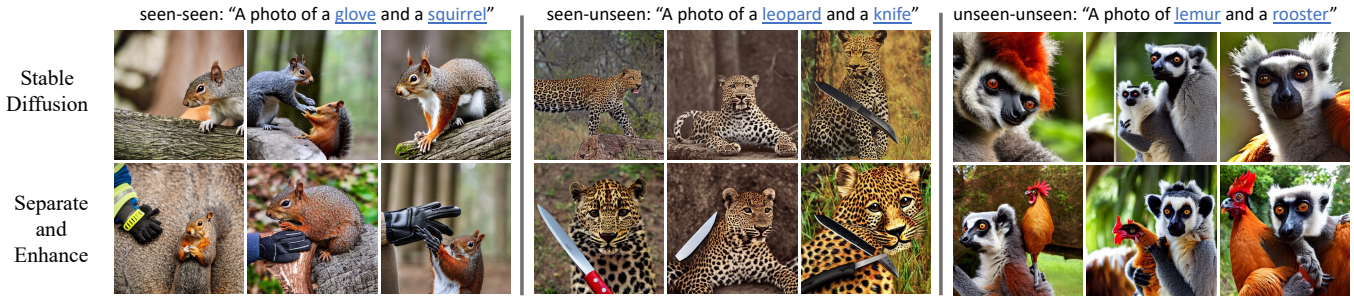

Figure 8: Visualizations for the large-scale experiments. Our model generates better text-aligned images with high quality for all three evaluation settings, indicating that our SepEn method has the capacity to jointly optimize a large collection of concepts with great generalizability to unseen novel concepts.

Table 2: Quantitative results for our method and the SD baseline. Ours significantly outperforms SD under all three regimes, indicating the great scalability of our model to a large collection of concepts and promising generalizability to novel concepts. By joint training with a large collection of concepts, our method better models the relationships between multiple concepts, thereby being able to tackle the more challenging fine-grained concepts in the large-scale setup.

Method	seen-seen		seen-unseen		unseen-unseen	
	Average Sim. Score (↑)	Success Rate (↑)	Average Sim. Score (↑)	Success Rate (↑)	Average Sim. Score (↑)	Success Rate (↑)
StableDiffusion	0.641 ±0.107	0.212	0.640 ± 0.105	0.227	0.633 ± 0.098	0.203
+ SepEn	**0.686 ± 0.107**	**0.299**	**0.677 ± 0.111**	**0.305**	0.679 ± 0.102	**0.294**
+ SepEn*	0.681 ± 0.095	0.294	0.673 ± 0.093	0.291	**0.687 ± 0.099**	0.287

Table 3: User study with 38 respondents. We have the highest human preference, showing the effectiveness of our method.

Method	SD	A & E	A-Star	SepEn (Ours)
Human preference	3.9%	13.5%	25.6%	53.9%

Table 4: Ablation study with $\mathcal{L}_{\text{Sep}}$ and $\mathcal{L}_{\text{En}}$. The Enhance loss holds better results in improving the text-image alignment while the Separate loss is more effective in maintaining the realism. Balancing them yields the best performance.

Method	FID (↓)	Average Similarity Score (↑)	Success Rate (↑)
StableDiffusion	32.96	0.742±0.091	0.209
+ $\mathcal{L}_{\text{Sep}}$	**36.33**	0.761±0.080	0.363
+ $\mathcal{L}_{\text{En}}$	42.84	0.770±0.093	0.374
+ SepEn	36.85	**0.809 ± 0.086**	**0.410**

SD is renowned for its capacity to generate high-quality images, and one of our objectives is to preserve this image fidelity post-finetuning, we use generated images for *single-object* prompts from SD as the source images. Notice that FID is not applicable to the Stable Diffusion baseline, mainly due to its tendency to underperform in generating images with multiple objects, despite generating photorealistic single-object images. However, we still report its FID for single-concept evaluation in Table 1 for reference. (2) The **Average text-text Similarity Score** proposed by Chefer et al. [2023] measures if the generated contents match the input prompts. For each prompt, we compute the average BLIP [Li et al. 2022] cosine similarity between the text prompt and the corresponding set of generated images. (3) The **Success Rate** measures if the output images contain all objects mentioned in the text prompt. Specifically, we use a pretrained detection model [Zhou et al. 2022] on ImageNet-21K to detect all possible objects from the given prompt. We count as a success case if the highest confidence scores for all target objects are larger than 0.7.

Implementation Details. We use SD 1.4 as the pretrained model for a fair comparison with Chefer et al. [2023] and Agarwal et al. [2023]. We set λ_E to 1.0 and λ_E to 2.0. λ_N is set to 0 and 0.5 for individual and large-scale training. We finetune our method for 200 steps for individual concepts and 10,000 steps for the large-scale experiments. More details are included in the supplementary.

4.1 Individual Prompt Evaluation

Additional setup. As the two baselines [Agarwal et al. 2023; Chefer et al. 2023] refine the latent outputs via test-time adaptation (TTA), we additionally report the quantitative results of a variant of our method, *SepEn (TTA)*, following the same manner as the two baselines. It applies the two proposed objectives to refine the latent at each time step during inference and can be referred to as the performance upper bound for our model.

Comparisons with baselines. We show the qualitative and quantitative comparisons with baselines in Figure 7 and Table 1 respectively. More visualizations are included in the supplementary. We observed that: (1) Stable diffusion usually generates photo-realistic yet misaligned single-object images, demonstrating the common issue of lacking compositional capacity. Compared with SD, our approach is able to generate images significantly better aligned with the input prompts while keeping good realism at the same time. (2) For the A & E baseline, though it can generate images better aligned with the input prompts, the visual quality has dropped a lot. We outperform this model under both realism and text-image alignment, owing to the two proposed novel objectives. (3) Even under the finetuning setting, we already outperform most of the TTA-based baseline models under all the metrics. Under the test-time adaptation setting, our model outperforms all the baselines, including the most recent A-Star, for the text-image alignment. However, its visual quality has dropped compared with our finetuned model.

Table 5: Quantitative comparison with finetuning different parameters. Updating *to_v* hurts the performance. Compared with *to_q*, tweaking *to_k* is the optimal choice.

Method	FID (↓)	Average Similarity Score (↑)	Success Rate (↑)
StableDiffusion	32.96	0.742±0.091	0.209
+ *to_q*	39.71	0.781 ± 0.102	0.365
+ *to_q* + *to_k*	51.60	0.759 ± 0.099	0.347
+ *to_k* + *to_v*	445.01	0.213 ± 0.051	0.004
+ *to_k* (ours)	**36.85**	**0.809 ± 0.086**	**0.410**

User study. Finally, we conduct a user study among SD [Rombach et al. 2022], A & E [Chefer et al. 2023], A-Star [Agarwal et al. 2023], and ours. Each subject is asked to choose the most preferred method with the best text-image alignment. The user preference results obtained from 38 subjects with 304 votes are shown in Table 3. Among all these methods, our results have the highest human preference, demonstrating the effectiveness of our approach.

4.2 Large-Scale Experiments

Additional setup. We propose three different types of evaluations: (1) **seen-seen** evaluation contains 100 pairs of concepts randomly sampled from the 200 training concepts. (2) **Seen-unseen** evaluation contains 80 pairs of concepts. For each of the concepts in the held-out category, we randomly select one concept in the training set. (3) **Unseen-unseen** evaluation contains 20 pairs of concepts. We randomly sample another concept from the held-out set for each unseen concept. We only compare with the Stable Diffusion baseline in this section, as the other baselines cannot generalize to a large scale of concepts. We additionally include a variant, SepEn* , which performs compositional finetuning on every single prompt. We report the Average Similarity Score and Success Rate. We also set the confidence threshold of the detector to 0.3, as some of the concepts are even very challenging for the detector.

The qualitative and quantitative results are reported in Figure 8 and Table 2 respectively. Firstly, Our model significantly outperforms the Stable Diffusion baseline both qualitatively and quantitatively, indicating the decent scalability of our model to a large collection of concepts, owing to our compositional finetuning strategy. More importantly, by jointly training with a large collection of concepts, our method learns a *global representation* of multiple concepts which better models the relationships between multiple concepts. Therefore, our large-scale model is able to (1) outperform the single-concept models and (2) generalize well to *novel* (unseen) concepts with the similar BLIP similarity score and success rate as the training ones, even for the most challenging unseen-unseen pairs. These results further convince our goal of **general compositional finetuning**. However, there is still a large room for improvement regarding the overall success rate, indicating the general challenge of the text-image alignment task of T2I models.

4.3 Ablation Study

Contributions of the two objectives. We ablate the contributions of the Separate loss $\mathcal{L}_{\text{Sep}}$ and the Enhance loss $\mathcal{L}_{\text{En}}$ in Table 4. We find that $\mathcal{L}_{\text{En}}$ holds better results in improving the text-image compositional alignment while $\mathcal{L}_{\text{Sep}}$ is more effective in maintaining the realism of the generated images. Both objectives can bring

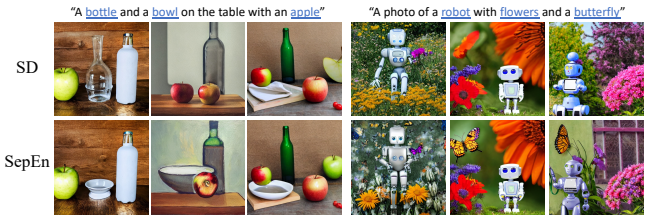

Figure 9: Extension of our method to more than two concepts. Our method is able to synthesize high-quality images matching the input prompts, indicating good generalizability.

improvements to the compositional capacity of diffusion-based T2I models, and balancing them yields state-of-the-art performance.

Finetuning different parameter groups. In Section 3, we posit the key role of finetuning the key mapping (*to_k*) functions for cross-attention modules. To validate our design, we further build 3 additional variants that finetune (1) both key mapping and value mapping functions (*to_v*); (2) both key mapping and query mapping (*to_q*) functions; and (3) only query mapping functions. Table 5 shows that tuning the value mapping functions leads to a decreased performance, verifying the role of the value mapping function in forming the object representations which should be kept frozen. Moreover, the query mapping function plays a less important role in the compositional synthesis task, involving or ignoring it has few differences while our original model design without tuning the query mapping function is effective and efficient.

Extension to more than two concepts. In Figure 9, we show the generalization of our algorithm to more than two concepts. We are still able to synthesize high-quality images matching the text prompts, indicating the decent generalizability of our method.

5 LIMITATION AND CONCLUSION

After large-scale finetuning, our model fails to distinguish the meaning of polysemy words (refer to Figure 10). We believe that employing a more advanced language model such as LLaMA [Touvron et al. 2023], coupled with a more diverse training process, could assist in addressing this limitation.

In summary, we address the compositional misalignment issue of diffusion-based T2I models. We propose a compositional finetuning strategy, *Separate-and-Enhance*, by incorporating two novel loss functions. Through extensive experimental evaluations, we demonstrate that our method design is promising for enhancing the compositional capacity of diffusion models. We also showcase that the proposed method is generally helpful for a large collection of concepts and achieves a great generalization capacity.

We include visualizations for failure cases (Figure 10), comparisons between SepEn and the individually optimized models SepEn* (Figure 11), large-scale experiments (Figure 12), and single-prompt evaluations (Figure 13) in the end.

ACKNOWLEDGMENTS

We thank Ziqi Pang, Deepak Pathak, Cherry Zhao, Jun-Yan Zhu, and Zhen Zhu for helpful discussion and feedback. This work was done when Zhipeng was an Adobe Research intern. This research is funded in part by Toyota Research Institute, NSF Grant 2106825, and NIFA Award 2020-67021-32799.

REFERENCES

Aishwarya Agarwal, Srikrishna Karanam, KJ Joseph, Apoorv Saxena, Koustava Goswami, and Balaji Vasan Srinivasan. 2023. A-STAR: Test-time Attention Segregation and Retention for Text-to-image Synthesis. In *ICCV*.

Jean Babaud, Andrew P Witkin, Michel Baudin, and Richard O Duda. 1986. Uniqueness of the Gaussian kernel for scale-space filtering. *PAMI* (1986).

David Bau, Alex Andonian, Audrey Cui, YeonHwan Park, Ali Jahanian, Aude Oliva, and Antonio Torralba. 2021. Paint by word. *arXiv preprint arXiv:2103.10951* (2021).

Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. 2023. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. In *SIGGRAPH*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

DC Dowson and BV Landau. 1982. The Fréchet distance between multivariate normal distributions. *Journal of multivariate analysis* (1982).

Dave Epstein, Allan Jabri, Ben Poole, Alexei A Efros, and Aleksander Holynski. 2023. Diffusion self-guidance for controllable image generation. In *NeurIPS*.

Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. 2023. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *ICLR*.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *NeurIPS* (2014).

Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. 2022. Vector quantized diffusion model for text-to-image synthesis. In *CVPR*.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626* (2022).

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *NeurIPS* (2020).

Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. 2023. Improving sample quality of diffusion models using self-attention guidance. In *ICCV*.

Ziqi Huang, Tianxing Wu, Yuming Jiang, Kelvin CK Chan, and Ziwei Liu. 2023. ReVersion: Diffusion-Based Relation Inversion from Images. *arXiv preprint arXiv:2303.13495* (2023).

Yunji Kim, Jiyoung Lee, Jin-Hwa Kim, Jung-Woo Ha, and Jun-Yan Zhu. 2023. Dense text-to-image generation with attention modulation. In *ICCV*.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2023. Multi-Concept Customization of Text-to-Image Diffusion. In *CVPR*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*.

Yumeng Li, Margret Keuper, Dan Zhang, and Anna Khoreva. 2023. Divide & Bind Your Attention for Improved Generative Semantic Nursing. In *BMVC*.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. 2022. Compositional visual generation with composable diffusion models. In *ECCV*.

Wan-Duo Kurt Ma, JP Lewis, W Bastiaan Kleijn, and Thomas Leung. 2023. Directed Diffusion: Direct Control of Object Placement through Attention Guidance. *arXiv preprint arXiv:2302.13153* (2023).

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741* (2021).

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. 2022. On Aliased Resizing and Surprising Subtleties in GAN Evaluation. In *CVPR*.

Pietro Perona and Jitendra Malik. 1990. Scale-space and edge detection using anisotropic diffusion. *PAMI* (1990).

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* (2022).

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *ICML*.

Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. 2023. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. *NeurIPS* (2023).

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. In *ICML*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402* (2022).

Jaskirat Singh and Liang Zheng. 2023. Divide, Evaluate, and Refine: Evaluating and Improving Text-to-Image Alignment with Iterative VQA Feedback. In *NeurIPS*.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2021. Score-based generative modeling through stochastic differential equations. In *ICLR*.

Michael Tomasello. 2009. *The cultural origins of human cognition*. Harvard university press.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *NeurIPS* (2017).

Ruichen Wang, Zekang Chen, Chen Chen, Jian Ma, Haonan Lu, and Xiaodong Lin. 2024. Compositional text-to-image synthesis with attention map control of diffusion models. In *AAAI*.

Weilun Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Dong Chen, Lu Yuan, and Houqiang Li. 2022. Semantic image synthesis via diffusion models. *arXiv preprint arXiv:2207.00050* (2022).

Joachim Weickert. 1998. *Anisotropic diffusion in image processing*. Teubner Stuttgart.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *CVPR*.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789* (2022).

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*.

Lvmin Zhang and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *ICCV*.

Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. 2023. LayoutDiffusion: Controllable Diffusion Model for Layout-to-image Generation. In *CVPR*.

Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. 2022. Detecting twenty-thousand classes using image-level supervision. In *ECCV*.

Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. 2019. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *CVPR*.

Figure 10: Failure case: after large-scale finetuning, the model fails to distinguish the polysemy words, *e.g.*, digital mouse and animal mouse; color orange and fruit orange. A better language model and a more diverse training process could help.

Figure 11: Visual comparisons between our method and the variant that individually optimizes each prompt (SepEn*). The compositional finetuning makes it possible for our method to scale up to a large collection of concepts to learn a global representation for multiple concepts with diffusion models, thereby outperforming the variant individually optimized for each pair.

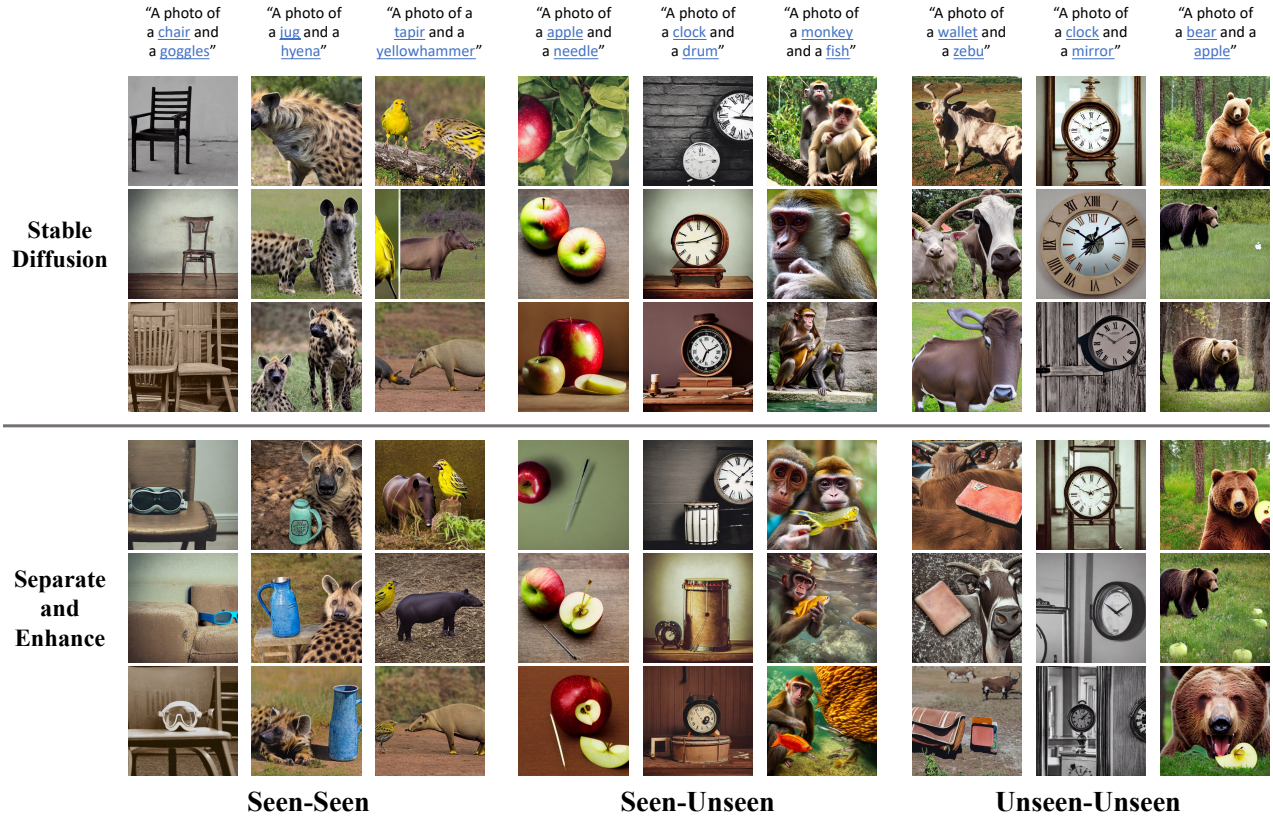

Figure 12: Visual comparisons for the large-scale experiments under the *Seen-Seen* (left), *Seen-Unseen* (middle), and *Unseen-Unseen* (right) settings. our method has better text-image alignments compared with the Stable Diffusion [Rombach et al. 2022] baseline while maintaining good visual quality.

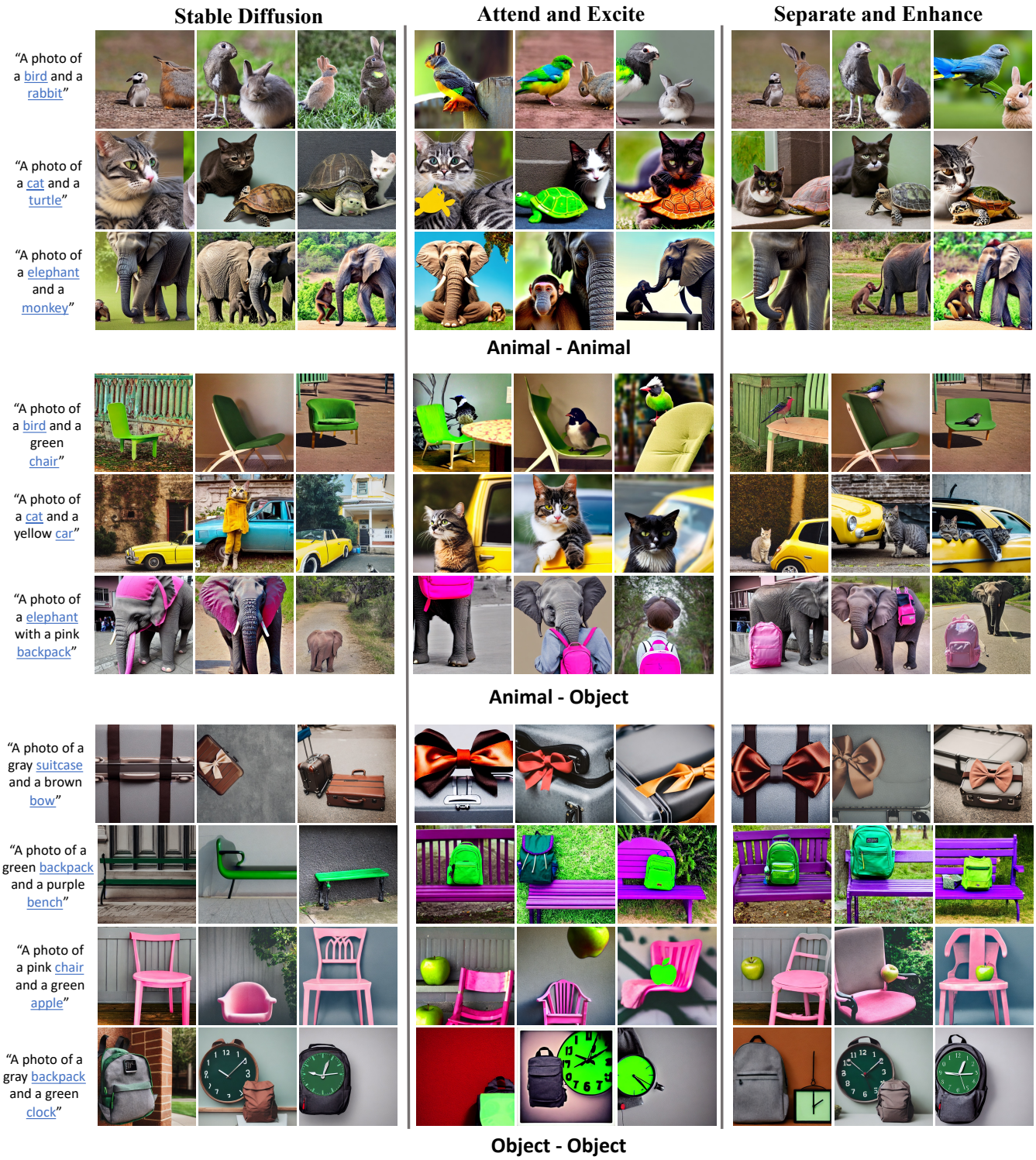

Figure 13: Visual comparisons for single-prompt evaluations for *animal-animal* (top), *object-object* (middle), and *object-object* (bottom) concept pairs. Our method outperforms the baselines under both realism and text-image alignment, owing to the two proposed novel objectives. It is noted that here we just finetune our model with a single-prompt for a fair comparison, but the goal of our approach is large-scale compositional finetuing with arbitrary prompts.