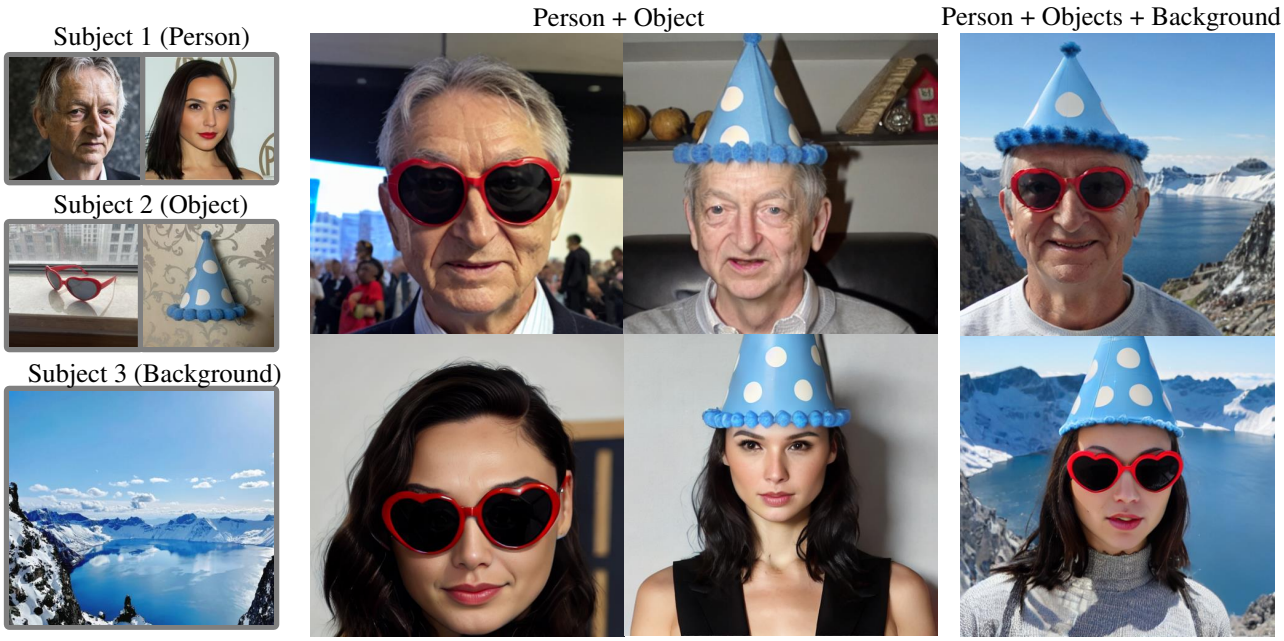
Cones: Concept Neurons in Diffusion Models for Customized Generation

Zhiheng Liu [1 * ‡] **Ruili Feng** [1 * ‡] **Kai Zhu** [1] **Yifei Zhang** [2 ‡] **Kecheng Zheng** [3] **Yu Liu** [4] **Deli Zhao** [4]
Jingren Zhou [4] **Yang Cao** [1]



(a) Reference images (b) Images generated by concatenating subject-specific concept neurons

Figure 1. We explore the subject-specific concept neurons in a pre-trained text-to-image diffusion model. Concatenating multiple clusters of concept neurons representing different persons, objects, and backgrounds can flexibly generate all related concepts in a single image.

Abstract

Human brains respond to semantic features of presented stimuli with different neurons. This raises the question of whether deep neural networks admit a similar behavior pattern. To investigate this phenomenon, this paper identifies a small cluster of neurons associated with a specific subject in a diffusion model. We call those neurons the concept neurons. They can be identified by statistics of network gradients to a stimulation connected with the given subject. The concept neurons demonstrate magnetic properties in interpreting and manipulating generation results. Shutting them can directly yield the related subject contextualized in different scenes. Concatenating multiple clusters of concept neurons can vividly generate all related concepts in a single image. Our method attains impressive performance for multi-subject customization, even four or more subjects. For large-scale applications, the concept neurons are environmentally friendly as we only need to store a sparse cluster of int index instead of dense float32 parameter values, reducing storage consumption by 90% compared with previous customized generation methods. Extensive qualitative and quantitative studies on diverse scenarios show the superiority of our

[*]Equal contribution . [‡]Work performed during internship at Alibaba DAMO Academy. [1]University of Science and Technology of China, Hefei, China [2]Shanghai Jiao Tong University, Shanghai, China [3]Ant Group, Hangzhou, China [4]Alibaba Group, Hangzhou, China. Correspondence to: Yang Cao <forrest@ustc.edu.cn>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

method in interpreting and manipulating diffusion models.

1. Introduction

The sophisticated structure of human brains allows miraculous cognitive and imaginative capabilities. Research has found that concept neurons in the human medial temporal lobe respond to semantic features of different presented stimuli separately (Bausch et al., 2021; Thiebaut de Schotten & Forkel, 2022). Those neurons encode temporal as well as abstract relations among elements of experience across spatiotemporal gaps and are thought to be the key to high-level intelligence (Bausch et al., 2021).

It is then curious to know, as one of the most successful artificial intelligence systems, do modern deep neural networks (LeCun et al., 2015) admit a similar structure of concept neurons. Specifically, to mimic the imaginative ability of the human brain, do generative diffusion models (Ho et al., 2020; Dhariwal & Nichol, 2021) encode different subjects separately with their neurons? This paper is about to answer this question from the perspective of subject-driven generation (Kumari et al., 2022; Ruiz et al., 2022). We propose to find a small cluster of neurons, which are parameters in the attention layer of a pretrained text-to-image diffusion model (Rombach et al., 2022), such that changing values of those neurons can generate a corresponding subject in different contents, based on the semantics in the input text prompt. We attribute these neurons as the concept neurons connected to the corresponding subject in the diffusion models. Finding them can advance our understanding of the underlying mechanism of deep diffusion networks and provide an original methodology for subject-driven generation.

We then study the interpretability of those concept neurons from several perspectives. We first investigate the robustness of concept neurons to changes in their values. We optimize a concept-implanting loss (Ruiz et al., 2022) on the concept neurons using float32, float16, quaternary, and binary (shutting those concept neurons directly without training) digital accuracy correspondingly. The results show similar performance among all the settings, demonstrating the strong robustness of concept neurons in controlling the target subject. While binary digital accuracy requires no further training and minimum storage space, we use it as our default subject-driven generation method. This method further admits fascinating additivity—concatenating concept neurons of multiple subjects directly can generate them all in the results, which may be the first to discover such a simple yet effective affine semantic structure in the parameter space of diffusion models. Further fine-tuning based on the concatenating can promote the multi-concept generation capability to a new milestone: we are the first to manage to generate four different diverse subjects in one image in the domain of subject-driven generation. Finally, thanks to their sparsity and robustness, the concept neurons can be efficiently used in large-scale applications. Storing the information to construct a given subject costs around only 10% of memory compared with previous subject-driven methods (Ruiz et al., 2022; Kumari et al., 2022), which is extremely economical and environment-friendly for commercial usage in mobile devices. Extensive studies on diverse categories, ranging from human portraits, scenes, decorations, *etc.*, demonstrate the superiority of over method in interpretability and multi-concept generation capability.

2. Preliminaries and Background

Diffusion Models. Diffusion models (Ho et al., 2020; Dhariwal & Nichol, 2021; Rombach et al., 2022; Song et al., 2020) are parametric neural networks that learn image distributions by gradual denoising. To further explore the extensibility of diffusion models, many works have been devoted to diffusion-based conditional generation, which can be broadly classified into two categories. The first one is the approach known as classifier-guidance (Liu et al., 2023), which utilizes a classifier to promote the sampling process of the pre-trained unconditional model. Despite the low cost, the generation effect is less competitive. The second one is known as the classifier-free approach (Ho & Salimans, 2022), which directly collects a large amount of data pairs for joint optimization under the guarantee of conditional probabilistic derivation. This approach can yield stunningly detailed results but requires a huge amount of data and computation resources. Owing to advances in language (Radford et al., 2021) and cross-modal foundation models (Radford et al., 2021), much text-to-image work (Saharia et al., 2022; Ramesh et al., 2022; Nichol et al., 2021) with classifier-free techniques is beginning to emerge, facilitating explicit control on the corresponding semantics and style. However, the expressiveness of text is still limited, and more work wants to utilize additional, conditional information (*e.g.*, reference image, grounding (Li et al., 2023) and sketch (Voynov et al., 2022)) to guide the global control further.

Text-to-Image Diffusion Model. A text-to-image diffusion model (Yu et al., 2022; Saharia et al., 2022; Rombach et al., 2022) $\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}$ will guide this denoising procedure with a text prompt describing the image content. Typically, it is trained by denoising a noised image $\boldsymbol{x}_t = \alpha_t \boldsymbol{x} + \sigma_t \boldsymbol{\epsilon}$ as

$$\mathbb{E}_{(\boldsymbol{x},\boldsymbol{c})\sim p_{\text{data}},t,\boldsymbol{\epsilon}}[\omega_t \|\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t, \boldsymbol{c}) - \boldsymbol{x}\|_2^2]. \quad (1)$$

Here $(\boldsymbol{x}, \boldsymbol{c})$ are (image, text prompts) pairs sampled from data; $\boldsymbol{\epsilon}$ are standard Gaussian noise added to the noised image; α_t, ω_t, and σ_t are hyper-parameter scalars to control the noise schedule evolved by time variable t from

$0, 1, \cdots, T$. After training, the model $\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}$ can generate various images described by the text prompts by denoising standard Gaussian noises. Throughout this work, we use Stable Diffusion V1.4 (Rombach et al., 2022) as the default text-to-image diffusion model due to its state-of-the-art performance and easy availability. However, Cones can also be simply applied to most text-to-image diffusion models like Imagen (Saharia et al., 2022) and DALLE-2 (Ramesh et al., 2022).

Customized Generation. The purpose of customized generation, as first proposed in DreamBooth (Ruiz et al., 2022), is to implant a given subject into the diffusion model and bind it with a unique text identifier to indicate its presence; so that the model can generate various renditions of the subject vividly guided by text prompts (Lu et al., 2020; Lee et al., 2019). To capture the subject $\mathcal{X}$, we need a few (usually 3 to 5) images of this subject $\{\mathcal{X}^i\}_{i=1}^s$ causally taken from different point-views and conditions. As in previous work (Ruiz et al., 2022), the subject $\mathcal{X}$ can be implanted to the diffusion model $\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}$ by minimizing a concept-preserving loss $\mathcal{L}_{\text{con}}$ together with a prior-preserving loss $\mathcal{L}_{\text{pr}}$ in the parameter space $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Let $\mathcal{X}_t^i = \alpha_t \mathcal{X}^i + \sigma_t \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathcal{O}, \mathcal{I})$, and $\boldsymbol{c}^{\mathcal{X}}$ be the text prompt 'A $\boldsymbol{V}^*$ [category name]' for this subject, then the subject-preserving loss is

$$\mathcal{L}_{\text{sub}} = \mathbb{E}_{\mathcal{X}^i, \boldsymbol{\epsilon}, t}[\omega_t \|\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\mathcal{X}_t^i, t, \boldsymbol{c}^{\mathcal{X}}) - \mathcal{X}^i\|_2^2]. \tag{2}$$

It explicitly binds the subject with the text identifier $\boldsymbol{V}^*$. To avoid over-fitting and language-drift (Ruiz et al., 2022), we further need the prior-preserving loss

$$\mathcal{L}_{\text{pr}} = \mathbb{E}_{(\boldsymbol{x}^{\text{pr}}, \boldsymbol{c}^{\text{pr}}), \boldsymbol{\epsilon}, t}[\omega_t \|\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t^{\text{pr}}, t, \boldsymbol{c}^{\text{pr}}) - \boldsymbol{x}^{\text{pr}}\|_2^2], \tag{3}$$

where $(\boldsymbol{x}^{\text{pr}}, \boldsymbol{c}^{\text{pr}})$ are image text prompt pairs with different subjects but the same category as the subject to implant. The full objective function is a combination of them as

$$\mathcal{L}_{\text{con}} = \mathcal{L}_{\text{sub}} + \lambda \mathcal{L}_{\text{pr}}. \tag{4}$$

3. Method

Our purpose here is to locate the corresponding neurons that control the generation of the given subject and use those neurons to guide customized generation (Ruiz et al., 2022; Kumari et al., 2022). For a diffusion model $\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}, \boldsymbol{\theta} = (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_n)^T \in \boldsymbol{\Theta} \subset \mathbb{R}^N$, where N is the parameter volume, we want to find a small collection of neurons $\boldsymbol{\theta}_{\mathcal{H}} = (\boldsymbol{\theta}_{h_1}, \cdots, \boldsymbol{\theta}_{h_n})^T, 1 \leq h_1 < h_2 < \cdots < h_n \leq N, n \ll N$, so that changes in them alone is enough to produce renditions of the given subject in different contexts, based on the text prompts.

This task significantly differs from previous customized generation work and is much more challenging. Here we not only pursue the generation quality of the subject but are also eager for the underline mechanism of how the diffusion model memorizes subjects in its parameter space. So our primary focus is on the interpretability of network neurons and how they influence the generation.

Advantages. Such methodology will allow significant practical advantages. As we will show in Sec. 3.1, simple statistics of network gradients can efficiently identify those concept neurons. Once locating them, we can *directly add* concept neurons of multiple subjects to generate them all together in the results, demonstrating the powerful interpretability of Cones as is discussed in Sec. 3.2. A couple of fine-tuning steps based on the above addition results can further enhance the generation in visual quality and multi-subject capability. To the best of our knowledge, this is the first method to manage to generate four different diverse subjects in one image. We will discuss this in detail in Sec. 3.3. Another superiority of Cones is its storage efficiency. Thanks to the sparsity and robust binary representation of concept neurons, storing concept neurons consumes only around 10% memory of Custom Diffusion (Kumari et al., 2022) and 0.05% memory of Dreembooth (Ruiz et al., 2022). This storage-friendly property enables large-scale commercial applications of customized generation.

3.1. Concept Neurons for a Given Subject

In this section, we analyze how neurons in a text-to-image diffusion model react to different subjects and locate the concept neurons corresponding to a given subject. While previous research shows that the K-V attention layer (Kumari et al., 2022) dominates the subject generation process, we follow its setting and limit the search for concept neurons in the parameters of the K-V attention layers. In what follows, we always assume that Θ is the parameter family for K-V attention layers.

Cones is inspired by Functional magnetic resonance imaging (fMRI) (Huettel et al., 2004) in medicine. It measures the small changes in blood flow that occur with brain activity. Research believes that different concepts, like visual objects or elements of experience, will induce blood-oxygen-level-dependent contrast (Logothetis et al., 2001; Kwong et al., 1992; Sharoh et al., 2019) in brain neurons of the human medial temporal lobe separately. Those brain neurons that prefer a specific concept are called brain concept neurons (Bausch et al., 2021) for this concept, and they dominate the brain response to this concept. We thus wonder whether there is also a preference for concept in the neurons of a pretrained diffusion model and whether they are responsible for generating this concept.

Specifically, we want to find a couple of neurons that, scaling down their absolute value, can reconstruct the

subject while maintaining prior information of the model, thus being able to generate the given subject in diverse contexts. This is equivalent to decreasing the value of the concept-implanting loss in Eq. (4). We do not consider the effect of scaling up for numerical reasons (values of scaling up can be up to infinity while scaling down is bounded by the initial values of neurons). Let $\theta = \boldsymbol{\theta}_h$ denote the h-element of the whole parameter vector $\boldsymbol{\theta}$ for simplicity. Scaling it by factor α will produce concept-implanting loss $\mathcal{L}_{\text{con}}(\alpha\theta)$. Let $\rho = (1-\alpha)(\theta\frac{\partial\mathcal{L}_{\text{con}}}{\partial\theta})^{-1}$. We can rewrite it as

$$\mathcal{L}_{\text{con}}(\alpha\theta) = \mathcal{L}_{\text{con}}(\theta(1-\rho\theta\frac{\partial\mathcal{L}_{\text{con}}}{\partial\theta})). \tag{5}$$

Through Taylor expansions (Rudin et al., 1976), we know that

$$\mathcal{L}_{\text{con}}(\alpha\theta) \approx \mathcal{L}_{\text{con}}(\theta) - \rho\theta^2\frac{\partial\mathcal{L}_{\text{con}}}{\partial\theta}^2 < \mathcal{L}_{\text{con}}(\theta) \tag{6}$$

as long as $0 < \rho \ll 1$. To make Eq. (5) is a scaling down, we need (when $0 < \rho \ll 1$)

$$0 < \alpha = 1 - \rho\theta\frac{\partial\mathcal{L}_{\text{con}}}{\partial\theta} < 1 \Leftrightarrow \theta\frac{\partial\mathcal{L}_{\text{con}}}{\partial\theta} > 0. \tag{7}$$

In conclusion, $\theta\frac{\mathcal{L}_{\text{con}}}{\partial\theta} > 0$ will indicate whether scaling down the h-th parameter will decrease the concept-implanting loss and thus identify whether θ_h is a concept neuron for the given subject $\mathcal{X}$. Rigorously, we can have the following theorem.

Theorem 3.1 (Identification of Concept Neurons)**.** *For a given parameter $\theta \in \boldsymbol{\theta}$, slightly scaling down it can decrease the concept-implanting loss, which is equivalent to*

$$\theta\frac{\partial\mathcal{L}_{\text{con}}}{\partial\theta} > 0, \tag{8}$$

and the decreasing value is proportional to $(\theta\frac{\partial\mathcal{L}_{\text{con}}}{\partial\theta})^2$. Thus θ is a concept neuron if and only if $\theta\frac{\partial\mathcal{L}_{\text{con}}}{\partial\theta} > 0$.

Following this theorem, a naive method to detect whether $\theta \in \boldsymbol{\theta}$ is a concept neuron can be that we sample K different values $\theta^1, \cdots, \theta^K$ ranging from zero to θ, and if

$$\theta^1\frac{\partial\mathcal{L}_{\text{con}}}{\partial\theta}(\theta^1) + \cdots \theta^K\frac{\partial\mathcal{L}_{\text{con}}}{\partial\theta}(\theta^K) > \tau > 0, \tag{9}$$

where $\frac{\partial\mathcal{L}_{\text{con}}}{\partial\theta}(\theta^k)$ is the gradient at point $\theta = \theta^i, k \in [K]$ and τ is a constant hyper-parameter, then θ is a concept neuron.

We deduce a self-adaptive sampling method for the choices of $\theta^1, \cdots, \theta^K$. We set $\theta^1 = \theta$, and

$$\theta^{k+1} = \theta^k(1 - \rho\theta^k\frac{\partial\mathcal{L}_{\text{con}}}{\partial\theta}(\theta^k)), k = 1, \cdots, K-1. \tag{10}$$

It will sample more densely in the neighborhood where $|\theta\frac{\partial\mathcal{L}_{\text{con}}}{\partial\theta}|$ is small, thus ambiguous to indicate the valuence, while sparsely when the valuence is obvious. Parameters with more ambiguous regions will tend to be excluded from concept neurons. Thus the identification of concept neurons will be more cautious and robust.

For all the parameters $\boldsymbol{\theta}$ of the diffusion model $\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}$, we can use Algorithm 1 to compute a concept neuron mask parallelistically to indicate whether each neuron is or not a concept neuron. The main computation Eq. (11) can be further accelerated using the Newton-Leibniz law of the calculus (Rudin et al., 1976). We use this accelerated version in practice. See appendix for detail.

Algorithm 1 Computing Concept Neuron Mask

Input: Concept-implanting loss function $\mathcal{L}_{\text{con}}$, parameter $\boldsymbol{\theta} \in \mathbb{R}^n$, maximum sampling number K, hyper-parameter $0 < \rho \ll 1$ and $\tau > 0$.
Set: $k = 1$ and $\boldsymbol{\theta}^k = \boldsymbol{\theta}$.
repeat
  compute

$$\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}^k \odot (\mathbf{1} - \rho\boldsymbol{\theta}^k \odot \nabla_{\boldsymbol{\theta}}\mathcal{L}_{\text{con}}(\boldsymbol{\theta}^k)); \tag{11}$$

  update $k = k + 1$;
until $k = K - 1$.
Compute: $\boldsymbol{M}_p = \boldsymbol{\theta}^1 \odot \nabla_{\boldsymbol{\theta}}\mathcal{L}_{\text{con}}(\boldsymbol{\theta}^1) + \cdots + \boldsymbol{\theta}^K \odot \nabla_{\boldsymbol{\theta}}\mathcal{L}_{\text{con}}(\boldsymbol{\theta}^K)$;
Set: $\boldsymbol{M} = \mathbf{1} - (\boldsymbol{M}_p > \tau)$.
Output: Binary concept neuron mask $\boldsymbol{M}$ to indicate whether each neuron is a concept neuron, 1 for not and 0 for is.

3.2. Interpretability of Concept Neurons

In this section, we explore various aspects of the interpretability of concept neurons. Based on our findings, we propose a new customized generation method, Cones, which implants the subject into a diffusion model by shutting the corresponding concept neurons.

Concept Neurons Indeed Responsible for Generation of the Corresponding Subject. In Fig. 2, which shows the attention map of the diffusion model $\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}$ before and after shutting the concept neurons corresponding to the given subject. We visualize the attention of each word in the text prompt. Shutting the concept neurons immediately draw the outline of the given subject in the attention map corresponding to the text identifier and subsequently generate the subject in the final output. This shows the strong connections between concept neurons and the given subject in the network representations.

Float32, Float16, Quaternary, and Binary Changes to the Concept Neurons Report Equal Effects. By our motivation, changes to concept neurons should decrease the

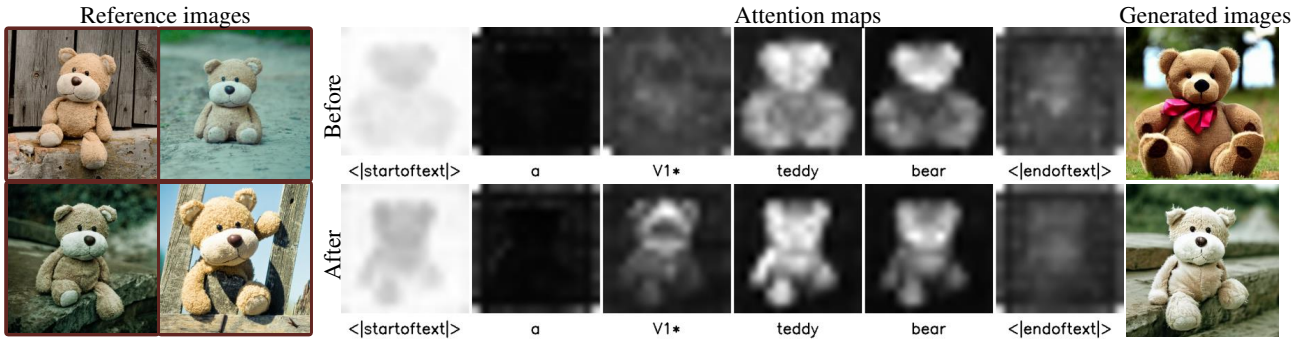


Figure 2. Attention maps before and after shutting subject-specific concept neurons. Shutting concept neurons draws the outline of given subject at the attention map of the text identifier $V1*$.

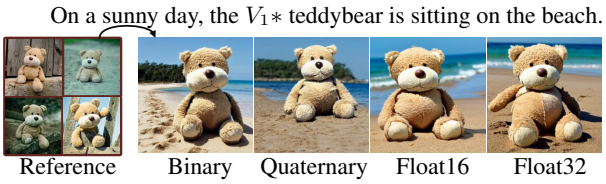


Figure 3. Results of optimizing the concept-implanting loss on concept neurons with float32, float16, quaternary, and binary digital accuracy. Binary digital accuracy corresponds to shutting the concept neuron without any further tuning. We can observe close performance for all cases. This demonstrates the strong robustness of controlling the subject generation of concept neurons.

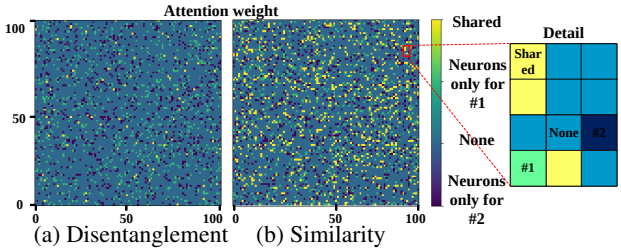


Figure 4. Disentanglement of concept neurons. (a) The intersection of concept neurons for two different subjects is sparsely distributed. (b) The concatenation of concept neurons for two different subjects is similar to the result of computing concept neurons from a joint loss of both subjects.

concept-implanting loss and thus generate the given subject. To demonstrate the strong interpretability and robustness of concept neurons, we study the effects of changes in different digital accuracy. We optimize the concept-implanting loss on the concept neurons and freeze the remains. We set the digital accuracy of the optimization to float32, float16, quaternary, and binary, in turn. We find close performance in all those cases, as is shown in Fig. 3. This demonstrates the strong robustness of concept neurons controlling the generation of the target subject.

Due to the strong performance and interpretability of binary digital accuracy, we name it ***Cones*** for customized generation and use it as our default method to implant the subject into diffusion models with concept neurons. Under this setting, we will first compute the concept neuron mask for a given subject and directly multiply it to the network parameter $\boldsymbol{\theta}$. The modified network $\hat{\boldsymbol{x}}_{M\odot\boldsymbol{\theta}}$ can then generate the target subject. This omits any further optimization on the concept neurons and is thus efficient and robust.

Additivity of Concept Neurons for Multi-Subject Generation. The concept neurons admit additivity. Direct joining concept neurons of multiple different subjects will yield concept neurons to generate the combination of them. This is the first time we manage to find an

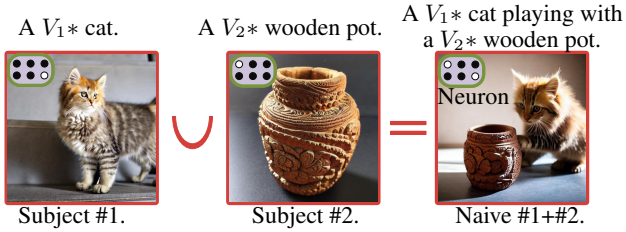


Figure 5. Additivity of concept neurons. Directly concatenating concept neurons of a subject cat and a subject wooden pot can vividly generate them both in the output under the direction of the text prompt. More results of additivity can be found in Sec. 4.

intrinsic affine semantic structure in the parameter space of diffusion models. In Fig. 5, we report the result of directly concatenating the concept neurons for a subject cat and wooden pot. After shutting the concatenated neurons, the diffusion model can immediately generate the two subjects together when accepting text prompts of the two corresponding identifiers as inputs. We will give more examples of the additivity for two and three concepts in Sec. 4. Directly concatenating concept neurons can be an efficient method for customized multi-subject generations.

3.3. Collaboratively Capturing Multiple Concepts

For better generation quality and multi-subject generating capability, we can further fine-tune the concept neurons after concatenating. We calculate the sum of concept-implanting losses for all the involved subjects as a multi-concept-implanting loss. We then replace $\mathcal{L}_{\text{con}}$ in Algorithm 1 with it and limit the computation in the concatenation of concept neurons for single subjects, as is shown in Fig. 6. This step will eliminate subtle conflicts in the concatenation of concept neurons due to inaccuracy in previous computations. The computed results can be more powerful in generating multi-subject. As we will show in Sec. 4, this is the first work to generate up to four different diverse subjects contextualized in one image.

Empirically we find the above pipeline slightly better than learning the concept neurons for multi-subject from scratch, *i.e.*, searching concept neurons in the whole K-V attention layers. This could be due to the increasing difficulty and instability of learning a loss landscape of complicated components. Besides, as we will show later, concept neurons enjoy good disentanglement; learning based on their concatenation could be stable and efficient.

Disentanglement of Concept Neurons. Concept neurons are well disentangled, which may be part of the reason for their additivity. Fig. 4 illustrates the concept neurons for the cat and wooden pot in Fig. 5 (in layer upblocks.2.attentions.1.transformerblocks.0.attn1.tov of the StableDiffusion V1.4). We can find that the shared neurons between two independent concepts are very sparse, counting merely 2.42% of the total concept neurons. We also report the differences between the concatenating two clusters of concept neurons and the result of learning from scratch. We can find the result of learning from scratch is close to the concatenation—they share 53.27% neurons. Thus when only two subjects are involved, learning based on concatenation performs similarly to learning from scratch. When involving more subjects, the increasing complexity of loss function often spoils learning from scratch. Learning from concatenation, on the other hand, provides a much more comfortable starting point and reduces the overall difficulty of the task.

3.4. Efficient Storage

Previous customized generation methods (Ruiz et al., 2022; Kumari et al., 2022)demand to save the parameters of the diffusion model in full digital accuracy. This can cost considerably for large-scale applications in mobile devices. While the concept neurons are sparse and binary, we only need to record a small collection of indices for them. Those indices can be stored with int instead of float data type, thus further reducing the storage consumption. As we will discuss in Sec. 4, Cones requires no more than 10% memory of previous customized generation methods.

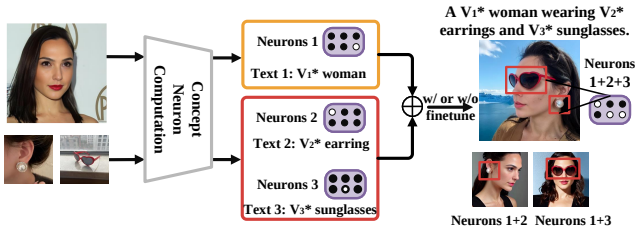


Figure 6. Illustration of collaborative capturing of multi-subject. Here we fine-tune the concatenation of concept neurons of multiple subjects to find a finer concept neuron mask.

4. Experiments

4.1. Implementation and Experiment Details

Evaluation Metrics, Datasets, and Implementation Details. We evaluate Cones with two famous metrics for customized generation proposed in Textual Inversion (Gal et al., 2022). (1) Image alignment, which measures the visual similarity between the generated images and the target concept. Specifically, we use the CLIP (Radford et al., 2021) model (ViT-L/14, consistent with the text encoder in Stable Diffusion V1.4) to calculate the CLIP-space cosine-similarity between the generated images and the target subject. For multi-subject generation, we calculate the image alignment of the generated images and each target subject separately and finally calculate the mean value. (2) Text alignment, which evaluates the ability of Cones to edit the target subjects with text prompts. To this end, we use a variety of prompts with different settings to generate images, including modifying the background, style, and attributes. We calculate the average CLIP-space embedding of the generated images and compute their cosine similarity with the CLIP-space embedding of the textual prompts, where we omit the text identifier in textual prompts. All images used in the paper are downloaded from anonymous e-commerce websites or Unsplash, like the dataset of Custom Diffusion (Kumari et al., 2022). Implementation details are reported in Appendix B.

Competing Methods. To evaluate our generation quality and multi-subject generation capability, we compare Cones with three competitors. They are Dreambooth (Ruiz et al., 2022) that fine-tunes all parameters in the diffusion model; Text Inversion (Gal et al., 2022) that adds a new token for each new concept and only updates the new token embedding during fine-tuning; and Custom Diffusion (Kumari et al., 2022) that optimizes the newly added token embedding in text encoder and a few parameters in diffusion model, namely the key and value mapping from text to latent features in the cross-attention (Yu et al., 2022; Vaswani et al., 2017) layers.

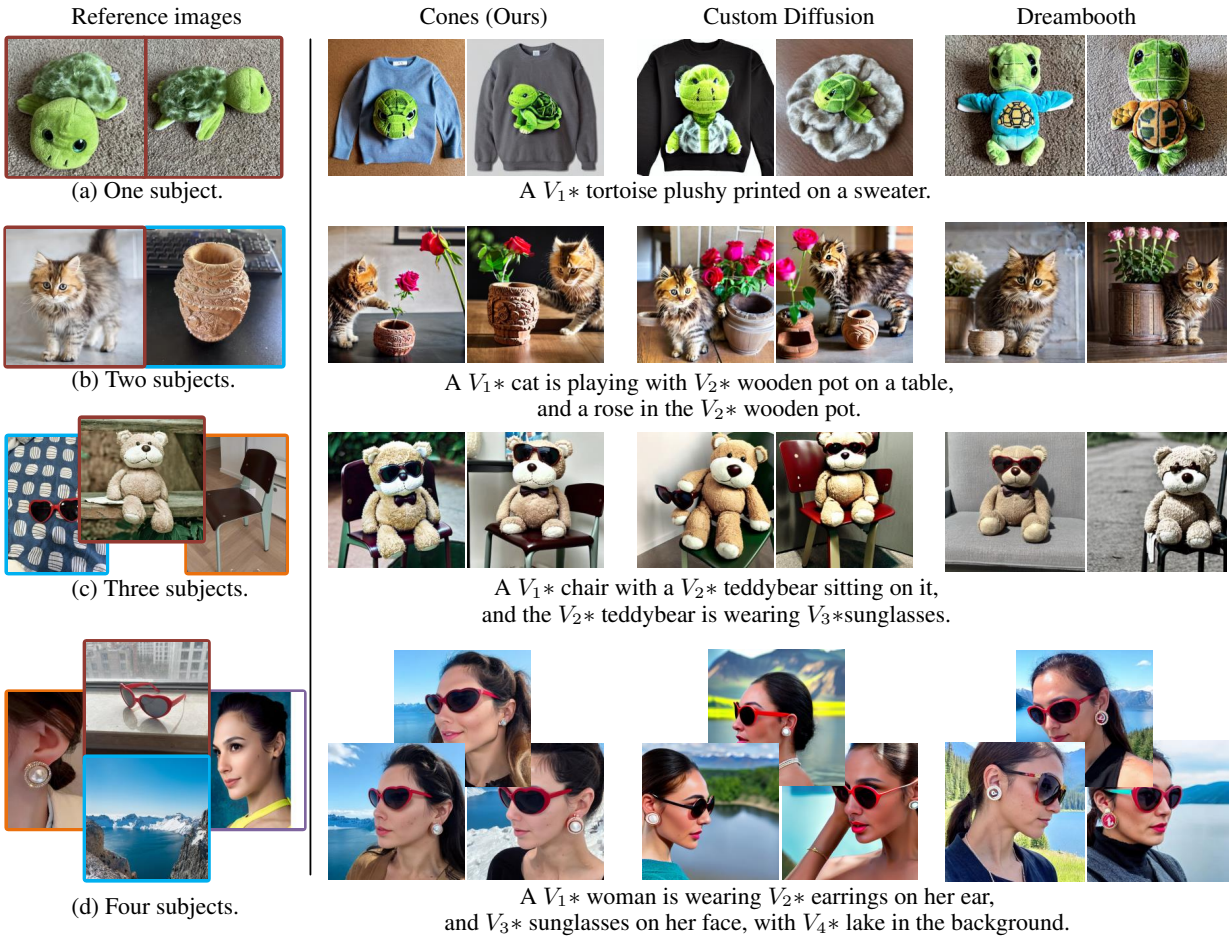


Figure 7. Comparison of multi-subject generation ability. First row: compared with other methods, ours can better generate the "sweater" in the prompt. Second row: Our method better reflects the semantics of "playing", while Dreambooth loses the details of the wooden pot. Third row: our generated images have a higher visual similarity with the target subject, and better semantics alignment with "sitting" and "wearing". Dreambooth fails to generate "chair". Fourth row: Cones (Ours) maintains high visual similarity for all subjects.

4.2. Qualitative Evaluation

To demonstrate the effectiveness of Cones, we conduct experiments on authentic images of diverse categories, including objects, backgrounds, portraits, *etc.*. As shown in Fig. 7, we show the results of generating several subjects in the same scene for the following four settings: (1) single subject: tortoise plushy, (2) two subjects: cat + wooden pot, (3) three subjects: chair + teddybear + sunglasses, (4) four subjects: woman + earrings + sunglasses + lake. For each method and subject setting, we sample 20 output images using 20 random seeds. We then select the best two of the 20 images as candidates for comparison. We omit Textual Inversion as it performs much less competitively. We thus put the generated results of Textual Inversion in Fig. A12. As shown in Fig. 7 (a), Cones is on-par with two other methods for visual similarity, but Cones has higher alignment with the input prompt than other methods for the single subject. As more subjects are composed together, Cones can generate images with good visual accuracy for all subjects. In contrast, the other two methods will make some subjects disappear or become less similar to the reference images.

Tuning-Free Comparison. By concatenating two clusters of concept neurons, we can realize the composition of the two concepts without further fine-tuning. While Custom Diffusion also provides a tuning-free method to composite multiple subjects (the "constraint optimization" method), we compare our concatenation of concept neurons with it in Fig. 8. It is easy to see that the concatenation of concept neurons significantly outperforms the tuning-free composition of the Custom Diffusion in visual quality and subject-generation accuracy.

4.3. Quantitative Evaluation and User Study

Quantitative Evaluation. We evaluate 20 prompts for each concept group and generate 50 images per prompt.

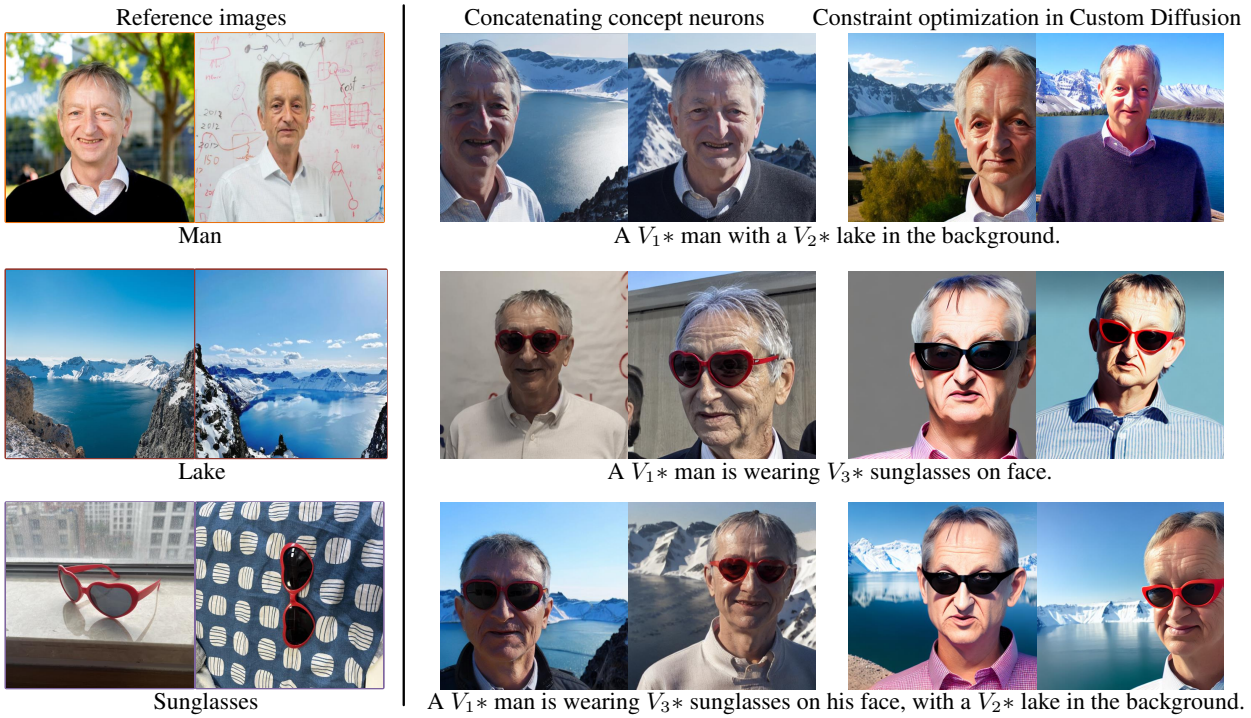


Figure 8. Comparison of tuning-free subject generation methods. For Cones, we concatenate concept neurons of multiple subjects directly. For Custom Diffusion, we use the "constraint optimization" method of it to composite multiple subjects.

For the multi-subject generation, we represent the image alignment as the mean of the visual similarity between the generated image and all target concepts. As shown in Tab. 1, For the single subject generation, our visual accuracy is comparable to Custom Diffusion, slightly lower than Dreambooth. Yet Cones has higher text alignment, which means Cones captures those subjects better and is more faithful to the prompt itself. When the number of involved subjects increases, Cones outperforms the competitors in all metrics.

User Study. We conduct a user study to further evaluate Cones. Two questions are designed to measure the image alignment and text alignment of all the methods. For text alignment, we ask the users "which image is most consistent with the textual description in the prompt". For image alignment, we ask the user "Which image is the most similar to the provided reference images". We hire 50 annotators to answer each of those questions. Details of how we conduct user study can be found in appendix. As shown in Tab. A3, Cones performs the best in all cases, earning the most votes, except for image alignment in the single subject generation.

	Method	Text-alignment	Image-alignment
Single-Subject	Textual Inversion	0.312	**0.744**
	DreamBooth	0.344	0.731
	Custom Diffusion	0.352	0.722
	Cones (Ours)	**0.361**	0.725
Two-Subjects	Textual Inversion	0.264	0.630
	DreamBooth	0.283	0.673
	Custom Diffusion	0.314	0.685
	Cones (Ours)	**0.337**	**0.698**
Three-Subjects	Textual Inversion	0.223	0.584
	DreamBooth	0.263	0.631
	Custom Diffusion	0.289	0.669
	Cones (Ours)	**0.301**	**0.685**
Four-Subjects	Textual Inversion	0.219	0.553
	DreamBooth	0.238	0.597
	Custom Diffusion	0.269	0.632
	Cones (Ours)	**0.285**	**0.653**

Table 1. Quantitative comparisons. Cones performs the best except for image alignment in the single subject case. This could be due to that the image alignment metric is easy to overfit as is pointed out in Custom Diffusion (Kumari et al., 2022). DreamBooth and Textual Inversion employ plenty of parameters in the learning, while Cones only involves the deactivation of a few parameters.

Sparsity and Storage. Thanks to the sparsity of concept neurons, we only need to record a small collection of indexes for them in attention layers. Those indices can be stored with int instead of float data type. As shown in Tab. 2, we show the storage required by Cones for multi-subject generation, and the sparsity of the corresponding concept neurons. Here sparsity means the percentage of concept neurons in all the neurons of the attention layers. We can

Method	Storage	Sparsity
Dreambooth	3.3GB	–
Custom Diffusion	72MB	–
Ours (single subject)	1.43MB$\pm$ 0.34MB	1.32% $\pm$ 0.29%
Ours (two subjects)	3.41MB $\pm$ 0.56MB	2.43% $\pm$ 0.44%
Ours (three subjects)	4.96MB $\pm$ 0.70MB	4.54% $\pm$ 0.59%
Ours (four subjects)	7.75MB$\pm$ 0.56MB	7.01% $\pm$ 0.26%

Table 2. Storage cost and sparsity of concept neurons. As the number of target subjects increases, we need to store more indexes of concept neurons. We save more than 90% of the storage space compared with Custom Diffusion,

find that Cones costs much less storage compared with the competitors.

5. Conclusion

This paper reveals concept neurons in the parameter space of diffusion models. We find that for a given subject, there is a small cluster of concept neurons that dominate the generation of this subject. Shutting them will yield renditions of the given subject in different contexts based on the text prompts. Concatenating them for different subjects can generate all the subjects in the results. Further fine-tuning can enhance the multi-subject generation capability, which is the first to manage to generate up to four different subjects in one image. Comparison with state-of-the-art competitors demonstrates the superiority of using concept neurons in visual quality, semantic alignment, multi-subject generation capability, and storage consumption.

Acknowledgments

This work is supported by National Key R&D Program of China under Grant 2020AAA0105701 and Alibaba Group through Alibaba Research Intern Program

References

Bausch, M., Niediek, J., Reber, T. P., Mackay, S., Boström, J., Elger, C. E., and Mormann, F. Concept neurons in the human medial temporal lobe flexibly represent abstract relations between concepts. *Nature communications*, 12 (1):6164, 2021.

Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., and Cohen-Or, D. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *arXiv preprint arXiv:2301.13826*, 2023.

Dhariwal, P. and Nichol, A. Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

Feng, W., He, X., Fu, T.-J., Jampani, V., Akula, A., Narayana, P., Basu, S., Wang, X. E., and Wang, W. Y. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*, 2022.

Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., and Cohen-Or, D. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Huettel, S. A., Song, A. W., McCarthy, G., et al. *Functional magnetic resonance imaging*, volume 1. Sinauer Associates Sunderland, 2004.

Kumari, N., Zhang, B., Zhang, R., Shechtman, E., and Zhu, J.-Y. Multi-concept customization of text-to-image diffusion. *arXiv preprint arXiv:2212.04488*, 2022.

Kwong, K. K., Belliveau, J. W., Chesler, D. A., Goldberg, I. E., Weisskoff, R. M., Poncelet, B. P., Kennedy, D. N., Hoppel, B. E., Cohen, M. S., and Turner, R. Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. *Proceedings of the National Academy of Sciences*, 89(12):5675–5679, 1992.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *nature*, 521(7553):436–444, 2015.

Lee, J., Cho, K., and Kiela, D. Countering language drift via visual grounding. *arXiv preprint arXiv:1909.04499*, 2019.

Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., and Lee, Y. J. Gligen: Open-set grounded text-to-image generation. *arXiv preprint arXiv:2301.07093*, 2023.

Liu, X., Park, D. H., Azadi, S., Zhang, G., Chopikyan, A., Hu, Y., Shi, H., Rohrbach, A., and Darrell, T. More control for free! image synthesis with semantic diffusion guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 289–299, 2023.

Logothetis, N. K., Pauls, J., Augath, M., Trinath, T., and Oeltermann, A. Neurophysiological investigation of the basis of the fMRI signal. *nature*, 412(6843):150–157, 2001.

Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., and Zhu, J. DPM-Solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. *arXiv preprint arXiv:2206.00927*, 2022.

Lu, Y., Singhal, S., Strub, F., Courville, A., and Pietquin, O. Countering language drift with seeded iterated learning. In *International Conference on Machine Learning*, pp. 6437–6447. PMLR, 2020.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. 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.

Rudin, W. et al. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1976.

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

Sharoh, D., Van Mourik, T., Bains, L. J., Segaert, K., Weber, K., Hagoort, P., and Norris, D. G. Laminar specific fMRI reveals directed interactions in distributed networks during language processing. *Proceedings of the National Academy of Sciences*, 116(42):21185–21190, 2019.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Thiebaut de Schotten, M. and Forkel, S. J. The emergent properties of the connected brain. *Science*, 378(6619): 505–510, 2022.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

von Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., and Wolf, T. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022.

Voynov, A., Aberman, K., and Cohen-Or, D. Sketch-guided text-to-image diffusion models. *arXiv preprint arXiv:2211.13752*, 2022.

Yu, J., Xu, Y., Koh, J. Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B. K., et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022.

Appendix

A. Proof

A.1. Proof to Theorem 3.1

This is easy to see from Eqs. (5) to (7).

A.2. Further Acceleration of Eq. (11)

Let $\gamma = \xi\theta$ and $\boldsymbol{\gamma} = \boldsymbol{\xi} \odot \boldsymbol{\theta}$. Using $\boldsymbol{\gamma}$ to replace $\boldsymbol{\theta}$ as the parameter of $\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}$, and independent variable of $\mathcal{L}_{\text{con}}$, then by Newton-Leibniz law of calculus, we have

$$\frac{\partial \mathcal{L}_{\text{con}}(\boldsymbol{\gamma})}{\partial \xi} = \frac{\partial \mathcal{L}_{\text{con}}(\gamma)}{\partial \gamma}\frac{\partial \gamma}{\partial \xi} = \theta \frac{\partial \mathcal{L}_{\text{con}}(\gamma)}{\partial \gamma}, \tag{A12}$$

$$\nabla_{\boldsymbol{\xi}} \mathcal{L}_{\text{con}}(\boldsymbol{\xi} \odot \boldsymbol{\theta}) = \boldsymbol{\theta} \odot \nabla_{\boldsymbol{\gamma}} \mathcal{L}_{\text{con}}(\boldsymbol{\gamma}) = \boldsymbol{\theta} \odot \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{con}}(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\gamma}}. \tag{A13}$$

Thus it is easy to see, the gradient descent over function

$$\mathcal{L}_{\text{con}}(\boldsymbol{\xi} \odot \boldsymbol{\theta}) \tag{A14}$$

will yield update rule

$$\boldsymbol{\xi}^{k+1} = \boldsymbol{\xi}^k - \beta \nabla_{\boldsymbol{\xi}} \mathcal{L}_{\text{con}}(\boldsymbol{\xi} \odot \boldsymbol{\theta}) = \boldsymbol{\xi}^k - \beta \boldsymbol{\theta} \odot \nabla_{\boldsymbol{\gamma}} \mathcal{L}_{\text{con}}(\boldsymbol{\gamma}^k), \tag{A15}$$

$$\boldsymbol{\gamma}^{k+1} = \boldsymbol{\xi}^{k+1} \odot \boldsymbol{\theta} = \boldsymbol{\xi}^k \odot \boldsymbol{\theta} - \beta \boldsymbol{\theta}^2 \odot \nabla_{\boldsymbol{\gamma}} \mathcal{L}_{\text{con}}(\boldsymbol{\gamma}^k) = \boldsymbol{\gamma}^k - \beta \boldsymbol{\theta}^2 \odot \nabla_{\boldsymbol{\gamma}} \mathcal{L}_{\text{con}}(\boldsymbol{\gamma}^k), \tag{A16}$$

where $\boldsymbol{\theta}^2 = \boldsymbol{\theta} \odot \boldsymbol{\theta}$. When learning rate β is small, $\boldsymbol{\xi}$ is initialized as $\mathbf{1}$, and iteration step k is not large, $\boldsymbol{\gamma}^k \approx \boldsymbol{\theta}$, thus

$$\boldsymbol{\gamma}^{k+1} \approx \boldsymbol{\gamma}^k - \beta (\boldsymbol{\gamma}^k)^2 \odot \nabla_{\boldsymbol{\gamma}} \mathcal{L}_{\text{con}}(\boldsymbol{\gamma}^k) = \boldsymbol{\gamma}^k \odot (\mathbf{1} - \beta \boldsymbol{\gamma}^k \odot \nabla_{\boldsymbol{\gamma}} \mathcal{L}_{\text{con}}(\boldsymbol{\gamma}^k)). \tag{A17}$$

Note that this is actually Eq. (11) and our sampling rule in Eq. (10).

Thus, we can accelerate the computation of neuron concepts by setting $\boldsymbol{\gamma} = \boldsymbol{\xi} \odot \boldsymbol{\theta}$, initializing ξ at $\mathbf{1}$, and conducting gradient descent on $\mathcal{L}_{\text{con}}(\boldsymbol{\xi} \odot \boldsymbol{\theta})$ with learning rate β and optimization variable $\boldsymbol{\xi}$. When

$$\boldsymbol{\xi}^K \approx 1 - \beta(\boldsymbol{\gamma}^1 \odot \nabla_{\boldsymbol{\gamma}} \mathcal{L}_{\text{con}}(\boldsymbol{\gamma}^1) + \boldsymbol{\gamma}^k \odot \nabla_{\boldsymbol{\gamma}} \mathcal{L}_{\text{con}}(\boldsymbol{\gamma}^k)), \tag{A18}$$

we have

$$\boldsymbol{M}_p = \frac{1}{\beta}(\mathbf{1} - \boldsymbol{\xi}^K). \tag{A19}$$

So the concept neuron mask can be computed as

$$\boldsymbol{M} = \mathbf{1} - (\boldsymbol{M}_p > \tau) = \mathbf{1} - (\boldsymbol{\xi}^K < \mathbf{1} - \beta\tau). \tag{A20}$$

Thus we can use Algorithm A2 to compute the concept neuron mask in practice.

B. Experiment Setups

We supplement the experimental Setups of each method in this section. Consistent with the Custom Diffusion (Kumari et al., 2022) setting, we use Stable Diffusion V1.4 as the pretrained model. For a fair comparison, we use 50 steps of DPM-Solver (Lu et al., 2022) sampler with a scale 7.5 for all above methods. All experiments are conducted using an A-100 GPU. For the three methods, except for Textual Inversion, training steps increase linearly as the number of involved subjects increases, and we initialize the identifier with the same rare occurring token as in Custom Diffusion.

Algorithm A2 Accelerated Computation of Concept Neuron Mask

Input: Concept-implanting loss function $\mathcal{L}_{\text{con}}$ with parameter $\boldsymbol{\theta} \in \mathbb{R}^n$ replaced by $\boldsymbol{\theta} = \boldsymbol{\xi} \odot \boldsymbol{\theta}$, training step K, learning rate $0 < \rho \ll 1$ and $\tau > 0$.
Execute: K-step gradient descent updates to variable $\boldsymbol{\xi}$ with loss function $\mathcal{L}_{\text{con}}(\boldsymbol{\xi} \odot \boldsymbol{\theta})$ and learning rate ρ.
Compute: $\boldsymbol{M}_p = \frac{1}{\rho}(\mathbf{1} - \boldsymbol{\xi})$.
Set: $\boldsymbol{M} = \mathbf{1} - (\boldsymbol{M}_p > \tau)$.
Output: Binary concept neuron mask $\boldsymbol{M}$ to indicate whether each neuron is a concept neuron, 1 for not and 0 for is.

	Textual Inversion		DreamBooth		Custom Diffusion		Ours	
	Text Alignment	Image Alignment	Text Alignment	Image Alignment	Text Alignment	Image Alignment	Text Alignment	Image Alignment
Single Subject	18.83%	26.17%	26.17%	**28.00**%	26.83%	20.67%	**28.17**%	25.17%
Two Subjects	14.83%	18.17%	25.33%	25.67%	29.00%	27.00%	**30.83**%	**29.17**%
Three Subjects	10.67%	12.00%	24.67%	23.50%	30.83%	30.16%	**34.17**%	**34.33**%
Four Subjects	8.83%	7.83%	20.17%	22.67%	33.17%	34.17%	**37.83**%	**35.33**%

Table A3. User study results. The value represents the percentage of users that think the image generated by the corresponding method is the best. The results show that our method is the most preferred by users for multi-subject generation, on both image and text alignment.

B.1. Textual Inversion

We train with the recommended[1] batch size of 4, a learning rate of 0.005 (scaled by batch size for an effective learning rate of 0.02) for 5,000 steps. The new token embedding is initialized with the category name. When some categories require multiple tokens to represent, we choose to use an approximation word to summarize the multiple tokens, such as replacing "wooden pot" with "pot".

B.2. Dreambooth

We use the third-party implementation of huggingface (von Platen et al., 2022) for Dreambooth[2]. Training is with a batch size of 1, learning rate 5×10^{-6}, and training steps of 800.

B.3. Custom Diffusion

We use the official implementation[3] for Custom Diffusion, which is consistent with paper, *i.e.*, the batch size is set to 4, training steps is set to 600 and the basic learning rate is 10^{-5} and scaled by batch size for an effective learning rate of 4×10^{-5}.

B.4. Cones (Ours)

Our experiments are conducted on an A-100 GPU with a batch size of 2. We use Algorithm A2 to find the concept neurons. The base learning rate is set to 3×10^{-5}. we further scale the base learning rate of 6×10^{-5} by the number of GPUs and the batch size.For the single-subject generation, the base learning rate is set to 2×10^{-5}, which can get better results. We train 1,000 steps for a single subject.

B.5. User Study

For one- to four-subject generation tasks, we design three different subject combinations for each task. This will yield 12 subject combinations in total. For each subject combination, we design four different text prompts to generate images. Each text prompt will be combined with 50 random seeds to generate 50 outputs. The best 2 are selected to represent the result of the corresponding text prompt. We conduct this procedure to all four methods, which results in 48 image octuples. Each octuple contains two best images generated by each method with each text prompt and subject combination. The results of user study can be found in Tab. A3.

C. More Results

C.1. Sequential Training Comparison

As shown in Fig. A9, we also evaluate Cones of sequential training on two subjects. Specifically, we optimize the concept-implanting loss for the second subject while shutting the corresponding concept neurons of the first subject. In the case of sequential training, we observe severe forgetting of the first concept for Custom Diffusion and DreamBooth, while Cones performs much better.

[1] https://github.com/rinongal/textual_inversion
[2] https://github.com/huggingface/diffusers/tree/main/examples/dreambooth
[3] https://github.com/adobe-research/custom-diffusion

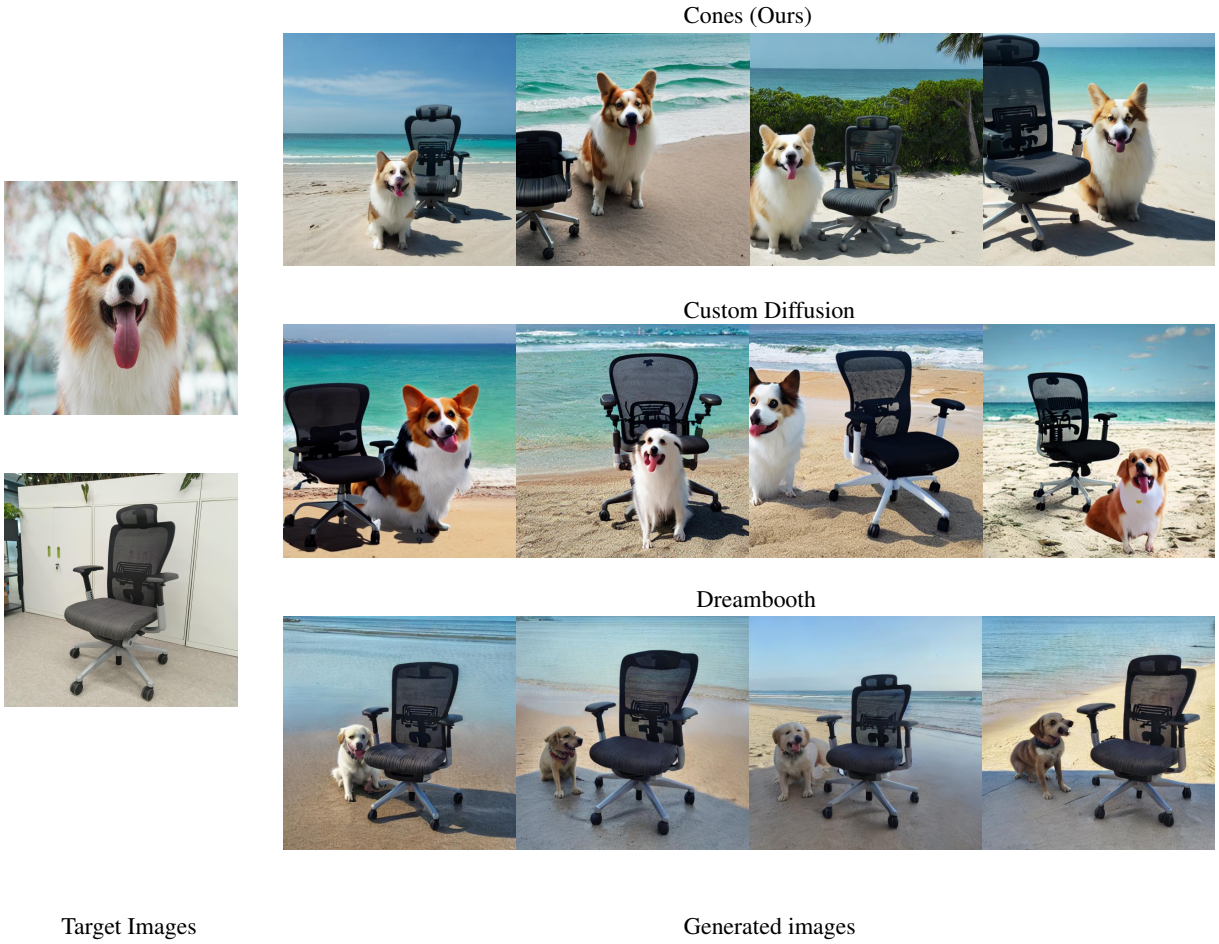


Figure A9. Sequential training results. The model learns ”dog” and ”chair” sequentially. It can be seen that the other two methods have severe forgetting of ”dog”. Cones retains better for both subjects.

C.2. Style Conversion

As shown in Fig. A10, Cones is able to express a certain style through text guidance and can also be fine-tuned on a fixed style.

C.3. Editing Performance

As shown in Fig. A11, Cones can capture more similarity in the generation results with the textual descriptions when editing images with text prompts, like expression switching and changing object colors, as well as adding a background.

C.4. Overfitting on the training prompt template

During fine-tuning, the target images are trained with the text prompt ”photo of a V_1* class”, where V_1* is the text identifier of the subject, and class is the class to which the subject belongs. As mentioned in Custom Diffusion (Kumari et al., 2022), after fine-tuning the models, the generations shift towards the target images and have less diversity compared to the pretrained model with the prompt ”photo of a V_1* class”. However, as shown in Fig. A13, the images generated by Cones have more diversity, which proves that Cones alleviates overfitting.

C.5. More results on multi subjects

As mentioned in Custom Diffusion (Kumari et al., 2022), the pretrained model encounters difficulty generating multiple subjects described in a single text prompt. As shown in Fig. A14, when Cones incorporates the Attend-and-Excite (Feng et al., 2022; Chefer et al., 2023) method to address this issue, it generates better results.

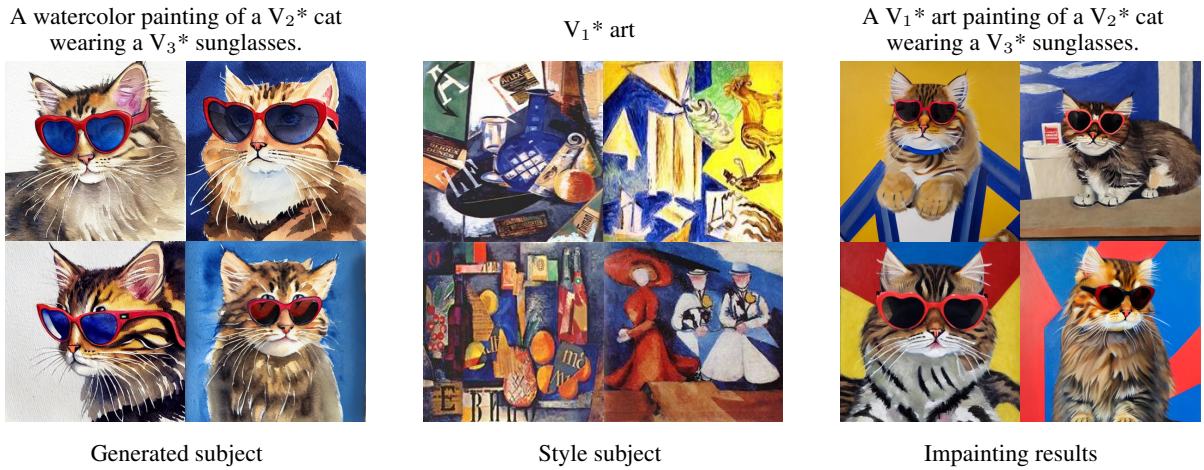


Figure A10. Style conversion results. The first column is the style conversion of the image through the knowledge of the pretrained model, the second column is a specific style, and the third column is the result of our generation in the specific style.

Figure A11. Editing results.

C.6. Interpolation results between various subjects

We conduct interpolation experiments in Fig. A15, where we can find a good semantic continuity among the interpolation points, and all intermediate points produce good visual results. This indicates that the concept neurons locate in the region of high semantic density and continuity.

C.7. Faliure modes

We show failure cases in Fig. A16, which may occur when juxtaposed displaying two subjects. This seems to be a common issue in Stable Diffusion as is pointed out in recent studies (Chefer et al., 2023; Feng et al., 2022). Also, more involved subjects usually decrease the success rate, we observe significantly more failure cases when generating five or more subjects.

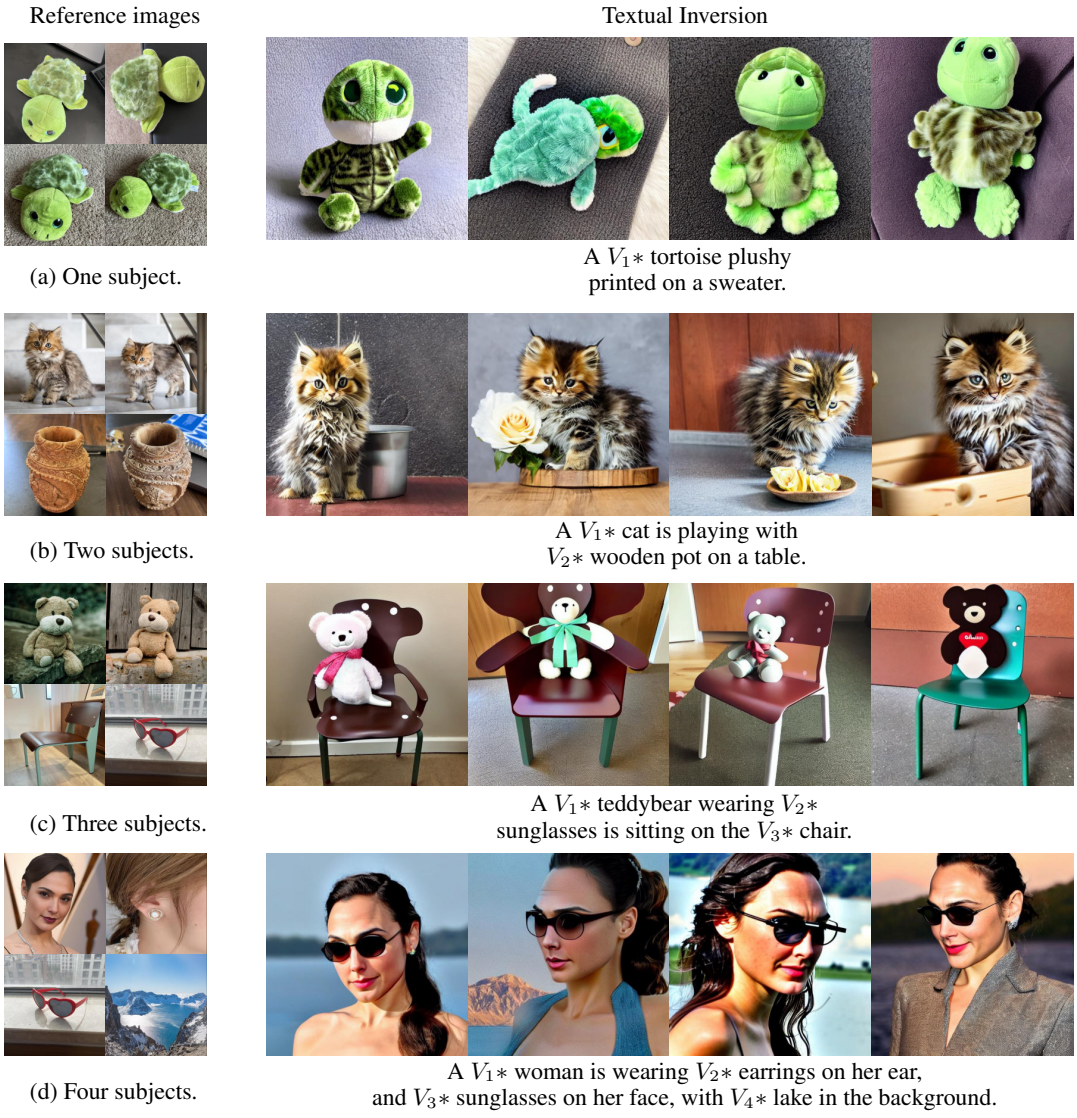


Figure A12. Multi-subject generation using Textual Inversion. We observe that Textual Inversion struggles with the composition of multiple subjects.

Figure A13. Overfitting on the training prompt template. The fourth column corresponds to the training data, and it can be seen that even without the text identifier, the generations of Custom Diffusion still retain some characteristics of the target images.The generations of Cones after finetuning has more diversity similar to the pretrained model.

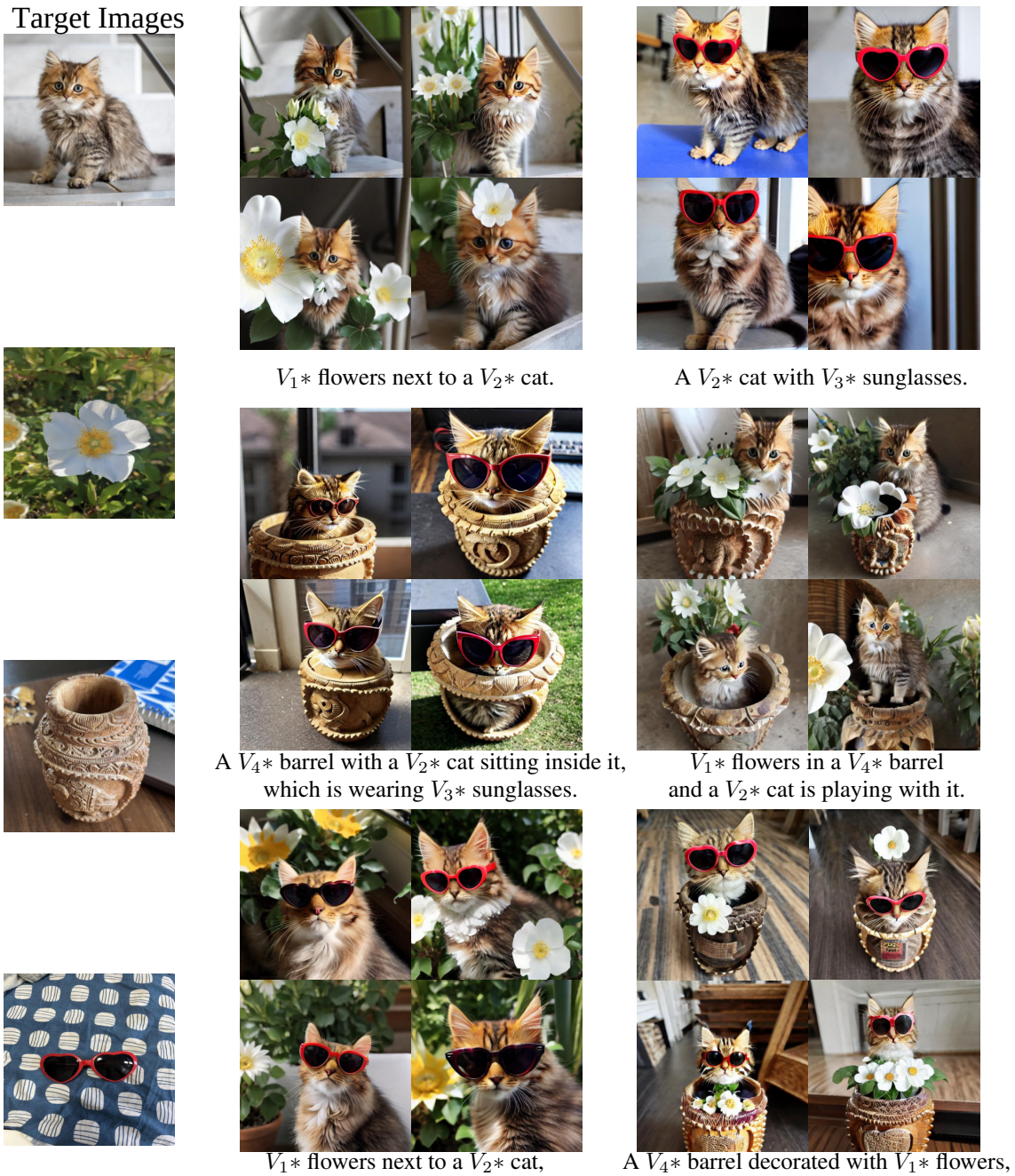


Figure A14. More results of multi-subject generation

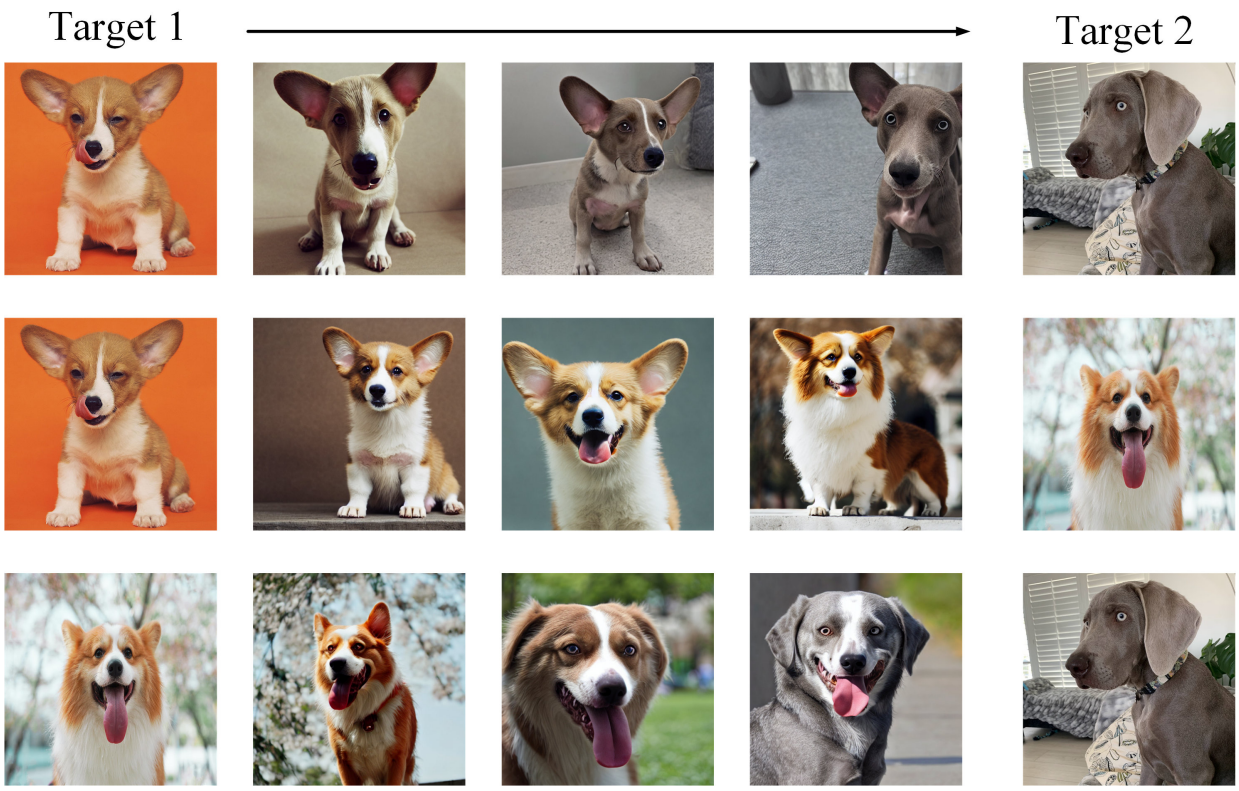


Figure A15. Interpolation results between the activation of various subjects. We generate images by interpolating the activation of the two subjects, the horizontal axis represents the different weights of the two subjects, and the middle image represents the equal weight of the two subjects.

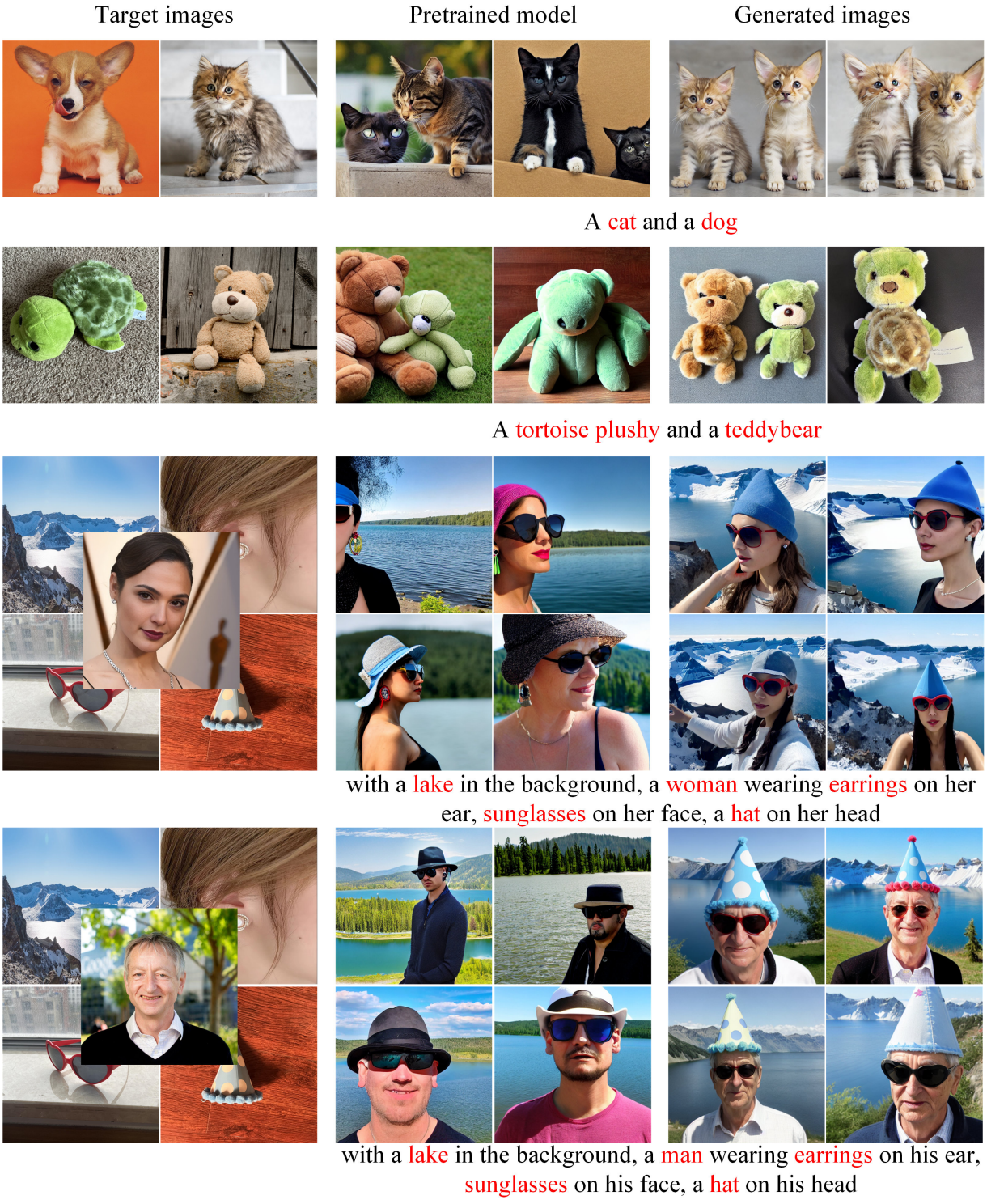


Figure A16. Failure cases.