Continual Personalization for Diffusion Models

Yu-Chien Liao*[1] Jr-Jen Chen*[1] Chi-Pin Huang[1] Ci-Siang Lin[1]
Meng-Lin Wu[2] Yu-Chiang Frank Wang[1]
[1]National Taiwan University [2]Qualcomm Technologies, Inc.

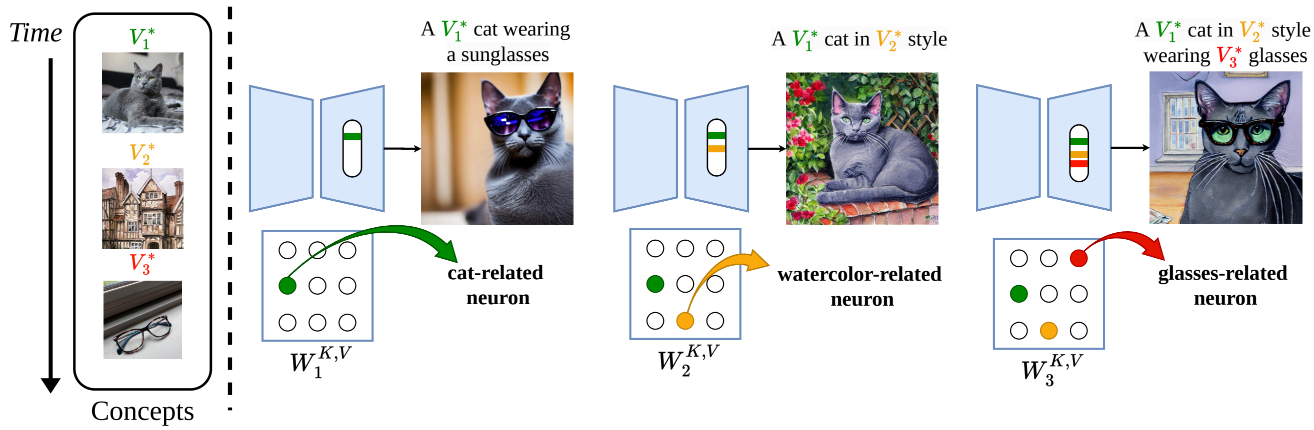

Figure 1. **Continual personalization.** We present **C**oncept **N**euron **S**election, CNS, a simple yet effective approach to incrementally customize visual concepts. By finetuning concept-related neurons, CNS preserves the zero-shot capabilities of pretrained diffusion models and alleviates catastrophic forgetting problems.

Abstract

*Updating diffusion models in an incremental setting would be practical in real-world applications yet computationally challenging. We present a novel learning strategy of **C**oncept **N**euron **S**election, a simple yet effective approach to perform personalization in a continual learning scheme.* CNS *uniquely identifies neurons in diffusion models that are closely related to the target concepts. In order to mitigate catastrophic forgetting problems while preserving zero-shot text-to-image generation ability,* CNS *finetunes concept neurons in an incremental manner and jointly preserves knowledge learned of previous concepts. Evaluation of real-world datasets demonstrates that* CNS *achieves state-of-the-art performance with minimal parameter adjustments, outperforming previous methods in both single and multi-concept personalization works.* CNS *also achieves fusion-free operation, reducing memory storage and processing time for continual personalization.*

* denotes co-first author.

1. Introduction

Latent Diffusion Models (LDMs) [29] represent a milestone for image generation task by leveraging vast collection of text-image pairs and denoising process. LDMs enable users to create high-quality images through simple text prompts. Yet, LDMs often fall short in generating user-specific concepts (*e.g.* their pets, a scene in a national park), which poses a practical challenge in text-to-image generation since these user-specific concepts are hard to describe directly by text. To address this issue, personalization techniques allow users to adapt LDMs to generate their desired specific content by finetuning it with their own examples. To realize single concept personalization, pioneer works first incorporate techniques such as prompt tuning [11] or weights finetuning [30]. However, when there are multiple concepts to be learned, naively applying these methods to compose multiple concepts in a single image usually results in overfitting and attributes binding [4, 10, 15, 20, 24, 39, 42], which means the model fail to correctly separate characteristics of each concept and generate mix-ups images. To address this concern, Kumari et al. [18] first aggregates multiple weights of personalized models by a constrained least square formula, which still results in huge performance degradation as the number of models' weights increase for multi-concept personalization. Many previous works [13, 17, 27, 32, 41] try to overcome this issue by reducing the optimized weights with LoRA [14] finetuning. A common approach nowadays is to learn each concept with LoRA weights separately and fuse the weights while generating multiple concepts images.

Existing methods [7, 13, 18, 19, 27] tend to make the assumption that personalized concepts are fixed, which means that storing all of the personalized model weights and multiple times of fusion are required if users need different numbers of personalized concepts across different images. However, in the realistic application scenario, users' personalized concepts never remain static and will incrementally increase. A practical scenario is that users can continuously assign new concepts to a single diffusion model and additional computation effort while generating new composite images is not required. Furthermore, learning the concepts separately usually results in conflict while fusing concepts together. Yang et al. [41] further prove that LoRA fusion methods will encounter concept vanishing and concept confusion even with additional information such as human poses or image layout.

In this paper, we propose CNS, **C**oncept **N**euron **S**election, which is able to identify the neurons relevant to personalized target concepts in an incremental fashion. Specifically, CNS allows one to automatically identify the neurons related to (few-shot) images of a concept (i.e., *base neurons*) and those related to the general image synthesis (i.e., *general neurons)* using diffusion models. By excluding general neurons from the base ones, the *concept neurons* describing the input concept of interest can be selected. Moreover, in order to achieve continual learning, an incremental finetuning scheme with such concept neurons is also proposed. Different from existing continual learning methods [8, 32], we do not require to train and store any extra LoRAs when handling multiple concepts. As a result, our CNS framework is able to achieve effective continual concept personalization, not only preventing the catastrophic forgetting but also preserving the zero-shot capability of pretrained text-to-image diffusion models.

Our contributions can be summarized as follows:

- We present CNS, which advances neuron selection with an incremental finetuning strategy for continual personalization.
- We introduce a unique process for *concept neurons* selection, which identifies neurons related to target concept images and distinguishes them from the irrelevant ones during continual learning.
- A incremental finetuning scheme is proposed with the selected *concept neurons*, which alleviates catastrophic forgetting while maintaining the zero-shot generation ability of the diffusion model.
- CNS is a fusion-free method for continual personalization. No additional model weights are needed to be stored, and no test-time optimization is needed either.

2. Related Work

2.1. Diffusion model personalization

Single and multiple-concept personalization. Concept personalization [5, 12, 16, 21, 26, 43] seeks to adapt pretrained diffusion models for synthesizing personalized concepts with only a few example images. Some common concepts are subjects, background or style. Previous works [11, 30, 37] achieved single concept personalization with several different strategies. For instance, Gal et al. [11] proposes to enables LDMs to learn personalized concepts through optimization of new textual embeddings. In the meantime, Ruiz et al. [30] assigns the user-specific concept to an unique identifier and finetunes the whole model. However, naively applying these methods to compose concepts in a single image usually results in overfitting and attributes binding. Kumari et al. [18] first proposes the joint-optimizing strategy and makes it realize to compose multi-concept in one image. Several following works [7, 13, 18, 19, 27] tried to make improvement in this field but still suffer from some common issues such as over-fitting on the target images and zero-shot ability of text-to-image degradation as mentioned in Sec. 1.

Low-Rank Adaption (LoRA) [14] finetuning shows the capability of achieving down-stream tasks while preventing overfitting. Thus, to mitigate the aforementioned overfitting issue to this topic, a common solution is learning new concepts with LoRA [14] finetuning. Previous works [13, 17, 25, 27, 31, 40, 41] propose different strategies to merge the learned LoRA weights to generate multi-concept images. For example, Gu et al. [13] proposes gradient fusion method, which extracts all input and output features from each personalized LoRA. Gradient fusion method aims to ensure that the output features of the fused model closely match the output features of all individual personalized LoRAs. However, since this method requires storing all input and output features during the fusion process, it demands significant memory space and computational time. Instead of fusing LoRAs after learning each concept, Po et al. [27] try to solve the problem from the initialization of LoRA weight. Under their assumption, they argue that as long as the down sampling part of LoRA weight across each concept are orthogonal to each other, then these LoRA weights keep independent and will not affect each other. However, as the number of concepts grows, the deviation of their assumption grows as well, which indicates the uncertainty of their method when fusing multiple personalized LoRA weights. Unfortunately, these aforementioned methods suffer from fusing multiple personalized LoRA, which also waste a lot of memory space and processing time.

Continual concept personalization. Continual concepts personalization further aims to incrementally extend the learned concepts in a single latent diffusion model. Smith et al. [32] propose a self-regularization loss between LoRA weights in different stage to preserve previous learned concepts. Nevertheless, as the number of concepts grows, this proposed regularization method lost its function since the new concept LoRA weight will face the restriction from all of the previous learned LoRA weights. Sun et al. [34] generate and store the data corresponding to the learned personalized concepts in a memory bank and wisely select images in the bank by their proposed rainbow-memory back strategy. Obviously, storing all of the relevant data of the learned personalized concepts is not realistic as the number of the concepts increases. Dong et al. [8] propose the task-shared project matrix to extract shared semantic information across different concepts and elastic weight aggregation during inference to tackle catastrophic forgetting. However, shared semantic information only exists between similar concepts, and elastic weight aggregation requires test-time optimization, which requires a lot of computational resources.

2.2. Neuron selection

As the number of parameters in large language models (LLMs) rapidly increases, how to efficiently finetune these LLMs has become a significant issue. Research [2, 6, 9, 33, 38] has provided evidence that certain neurons within feed-forward layers of transformer-based LLMs are intricately tied to task-specific outputs. Adjusting these neurons can significantly influence task performance, reducing the necessity for extensive finetuning. Building on this understanding, recent methods have been developed to detect modular structures within pretrained transformers, capitalizing on neuron sparsity [44]. Moreover, studies reveal that these modules possess specialized functions, each serving unique roles within the model [45].

After many works on LLMs prove the effectively of neuron selection and its widely usage, Liu et al. [22] first applied this concept in the text-to-image diffusion model generation and customization. However, [22] has been proved that suffering from high training costs and low success rates along with the increased number of subjects in [23], who needs an additional user-defined layout to improve it. Additionally, Chavhan et al. [3] builds upon this foundation, explores by identifying neurons responsible for generating undesired concepts within diffusion models. Unlike in LLMs, isolating such neurons in LDMs presents unique challenges due to the complex aggregation across multiple denoising steps and the model's sensitivity to intermediate outputs from prior time steps. Inspired by [3], we propose CNS to select sparse and highly related neurons in diffusion models to excel at the continual personalization.

3. Preliminary of Diffusion Models

Text-to-image diffusion models generate images by progressively removing noise from an initial noisy input, guided by a conditioning vector derived from a text prompt. These models integrate the conditioning vector through cross-attention layers, which adjust the network's latent features according to the text conditioning. This approach enables the model to produce outputs that accurately align with the prompt, allowing for precise control over the generated image's content. In cross-attention layers, let $\mathbf{c} \in \mathbb{R}^{s\times d}$ represent the text conditioning features with s as the number of tokens and d as the dimension of tokens. We also make $\mathbf{f} \in \mathbb{R}^{(h\times w)\times l}$ represent the latent image features. A single-head cross-attention operation [36] computes $Q = \mathbf{f}W^q$, $K = \mathbf{c}W^k$ and $V = \mathbf{c}W^v$, followed by a weighted sum:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d'}}\right)V, \quad (1)$$

where, W^q, W^k, and W^v project the inputs to the query, key, and value spaces, with d' as the dimension for key and query outputs. The attention block output is then used to update the latent features.

These models minimize a loss function that refines the denoising process to approximate the target image, typically formulated as follows:

$$\mathbb{E}_{\mathbf{x},\mathbf{c},\boldsymbol{\epsilon},t}\left[w_t\|\hat{\mathbf{x}}_\theta(\alpha_t\mathbf{x} + \sigma_t\boldsymbol{\epsilon}, \mathbf{c}) - \mathbf{x}\|_2^2\right], \quad (2)$$

where $\mathbf{x}$ is the ground-truth image, $\mathbf{c}$ is the conditioning vector obtained from the text prompt, and α_t, σ_t, w_t are parameters controlling the noise schedule at time t.

4. Method

4.1. Problem formulation and framework overview

In the continual learning scheme, we define the m representing the index of current concept to learn, and only N_m images associated with concept m is presented during training. Additionally, the model finetuned on m-th concept is expected to preserve the knowledge previously learned from concepts $1 : m-1$. In the following section, we omit m notation if we are discussing about only one concept.

We propose **C**oncept **N**euron **S**election, CNS, to achieve continual personalization while alleviating catastrophic forgetting. We introduce a neuron selection method that identifies a compact set of concept-specific neurons, which realizes the above continual learning scheme. As depicted in Fig. 2 and described in Sec. 4.2, our proposed method learns and distinguishes between neurons for concept personalization and image generation. As detailed in Sec. 4.3, only the neurons associated with the input concept need to be updated and regularized in the continual learning scheme. As confirmed in our experiments, our method

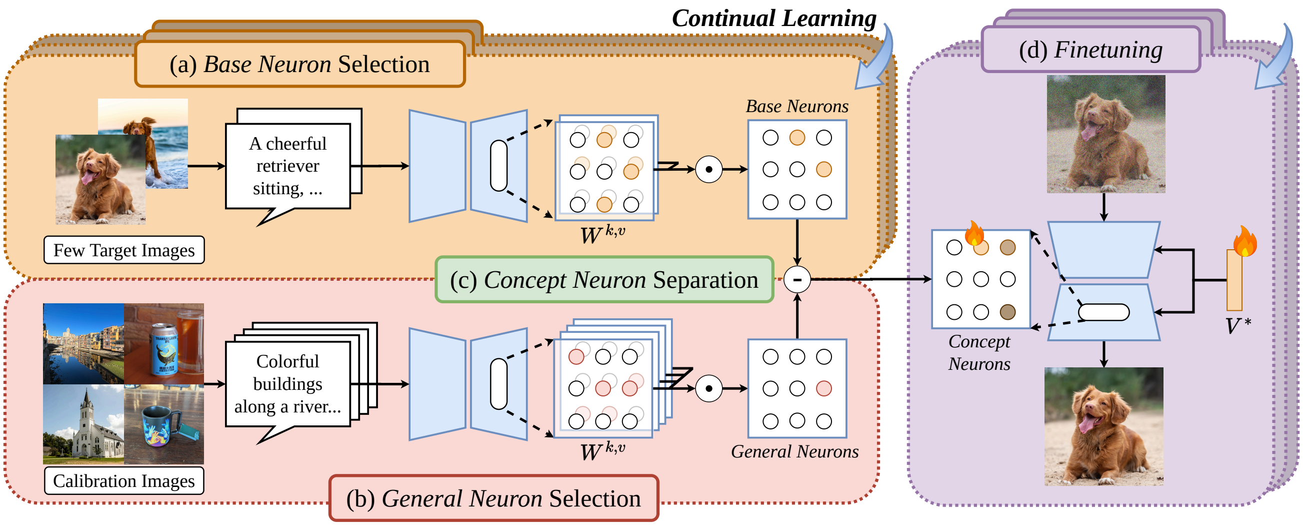

Figure 2. **Overview of CNS.** The proposed framework for neuron selection consists of (a) base neuron selection, (b) general neuron selection, and (c) concept neuron separation. With sparse, concept-specific neurons automatically selected fro each concept, the proposed incremental finetuning scheme in (d) update the text-to-image diffusion model for continual personalization.

clearly perform against existing approaches for single and multi-concept image personalization.

4.2. Learning of concept neurons

Base neuron selection. Recent works on concept editing and model pruning [3, 35] note that neurons with large responses to the objective imply higher contributions to the learned model, and thus they are preferred to be updated/edited during the training stage. Following the observation in [18] that the cross-attention layer parameters have relatively higher correlations to the image personalization objective, we thus focus on neurons of cross-attention layers in diffusion models for neuron selection.

To be more precise, we denote the weights of key and value mapping by $W^k \in \mathbb{R}^{d\times d_k}$ and $W^v \in \mathbb{R}^{d\times d_v}$, while the inputs of both mapping are text embedding denoted by $\mathbf{c}$. We obtain $\mathbf{c}$ text embedding from the concept image caption generated by a pretrained image captioning model [1]. To assess the significance of each element in both weights $W^{k,v}$for a single image, we calculate the element-wise product of its magnitude with the ℓ_2 norm of the text embedding feature $\mathbf{c}$ across dimensions following [35].

Take W^k as an example, we calculate the importance scores as follows:

$$\mathbf{S}(W^k, \mathbf{c}) = \left|W^k\right| \odot \left(\mathbf{1} \cdot \|\mathbf{c}\|_2\ \right), \tag{3}$$

where $|\cdot|$ represents the absolute value, $\|\mathbf{c}\|_2$ denotes the ℓ_2 norm applied to each column of $\mathbf{c}$, producing a vector of dimension d, and the symbol $\odot$ indicates element-wise matrix multiplication. Note that in Eq. (3), $\mathbf{1}\cdot\|\mathbf{c}\|_2$ specifically uses broadcasting to apply $\|\mathbf{c}\|_2$ across the rows of W, allowing for element-wise multiplication in each row. For a given row $W_{i,:}$ with the associated importance scores $\mathbf{S}(W,\mathbf{c})_{i,:} \in \mathbb{R}^{d\times d_k}$. We would select a neuron as a base one, if its score is above a pre-determined threshold (see details in the supplementary material).

Since there are N images for concept, the above process would select neurons for each image, which can be viewed as a binary neuron mask $\boldsymbol{M}_n, n \in \{1, 2, ..., N\}$. We then aggregate all neuron masks obtained across images of the same concept and apply a logical `and` operation. We refer to such aggregated neurons as , as depicted in Fig. 2(a). The resulting mask $\boldsymbol{M}^{base}$ can be expressed as:

$$\boldsymbol{M}^{base} = \bigwedge_n \boldsymbol{M}_n, n \in \{1, 2, ..., N\}. \tag{4}$$

Note that the above process is applied to both key and value mappings $W^{k,v}$ in all cross-attention layers.

General neuron selection. Despite the idea of selecting responsive neurons from concept images aligns with that of model pruning [3, 35], we observe a large overlap between neurons selected from the image diffusion model when different text prompts are served as the inputs. As shown in Fig. 3, by increasing the number of input prompts as detailed in supplementary, we empirically observe that about 53% of neurons are always selected. This suggests the aforementioned scheme not only selects neurons associated with the input prompt/condition but also identifies neurons contributing to the general image generation process. In CNS, we define the latter type of neurons as *gen-*

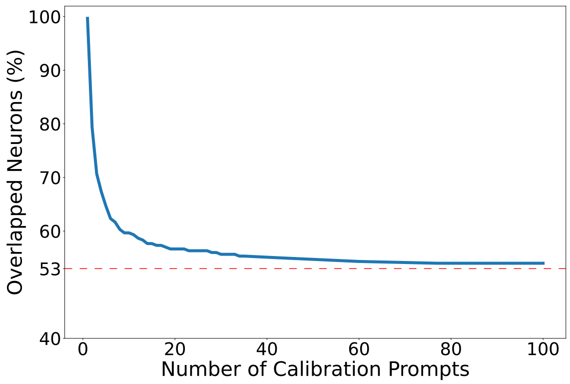

Figure 3. **Overlapped percentage of across images.** By increasing the number of text prompts, we observe a high percentage (around 53%) of shared across the resulting images. This suggests that a large portion of share the goal of image generation, not concept personalization.

eral neurons, and we aim to disregard such neurons when performing continual concept personalization.

We now discuss how we detect general neurons which are related to image generation but not describing the concept of interest. As depicted in Fig. 2(b), we collect a list of diverse calibration prompt set $\mathbf{P}_k$ for $k \in \{1, 2, \ldots, K\}$ to find out *general neurons* in latent diffusion models' cross-attention layers. We use $K = 20$ different prompts as calibration prompts to obtain the *general neurons* mask $\boldsymbol{M}^{general}$, which is obtained by:

$$\boldsymbol{M}^{general} = \bigwedge_k \boldsymbol{M}_k, k \in \{1, 2, ..., K\}. \tag{5}$$

Identification of concept neurons. With the collection of both and *general neurons*, it comes straightforward to exclude the *general neurons* from the for identifying the *concept neurons* of interest. This is simply achieved by performing a logical `not` operation on $\boldsymbol{M}^{base}$ and $\boldsymbol{M}^{general}$, resulting in the *concept neurons* mask $\boldsymbol{M}^{concept}$ for a specific concept. That is,

$$\boldsymbol{M}^{concept} = \boldsymbol{M}^{base} \wedge \neg \boldsymbol{M}^{general}, \tag{6}$$

where $\wedge$ and $\neg$ denote the logical `OR` and `NOT` operators. Again, we apply the aforementioned operation on both $W^{k,v}$ in all cross-attention layers. For each concept of interest, the derived *concept neurons* mask will be utilized for later continual learning. It is worth noting that even though we subtract general neurons from base one, those general neurons are fixed (not pruned) during training. Therefore, the associated pre-trained ability would not be affected.

4.3. Continual personalization with concept neurons

With *concept neurons* for each concept properly identified, the final challenge is to perform continual concept personalization while mitigating possible catastrophic forgetting problems. Since there is no guarantee that *concept neurons* of different concepts are distinct, we propose to perform continual learning with *neuron regularization*:

$$L_{reg} = \lambda_1 \| W_m \odot \boldsymbol{M}^{reg} - W_{m-1} \odot \boldsymbol{M}^{reg} \|_2 + \lambda_2 \| W_m \odot \boldsymbol{M}^{reg} - W_0 \odot \boldsymbol{M}^{reg} \|_2, \tag{7}$$

where $\boldsymbol{M}^{reg} = \boldsymbol{M}_m^{concept} \wedge (\bigvee_{m'=1}^{m-1} M_{m'}^{concept})$ denotes the regularization neuron masks, which indicates the intersection of the neurons to be updated and the neurons have been tuned for the previously learned concepts. W_m indicates the model weights of the neurons to be updated when learning the m^{th} concept, W_{m-1} indicates the model weights after customized on the $m-1^{\text{th}}$ concept and W_0 indicates the pretrained weights of the original diffusion model. In Eq. (7), the first term regularizes the model weights previously learned to prevent catastrophic forgetting (i.e., prior personalized concepts). On the other hand, the second term in Eq. (7) regularizes the model weights with the pretrained ones to prevent zero-shot capability degradation.

Combining with the loss function presented in [30] to prevent from subject overfitting, our proposed framework is capable of retaining both previously learned concepts from W_{m-1} and the inherent zero-shot text-to-image capability from the pretrained diffusion model W_0 during finetuning on new concept. In summary, the overall objective function of our proposed continual personalization loss via concept neurons selection is formulated as below:

$$\mathbb{E}_{\mathbf{x},\mathbf{c},\boldsymbol{\epsilon},\boldsymbol{\epsilon}',t}[w_t \| \hat{\mathbf{x}}_\theta(\alpha_t \mathbf{x} + \sigma_t \boldsymbol{\epsilon}, \mathbf{c}) - \mathbf{x} \|_2^2 + \lambda w_{t'} \| \hat{\mathbf{x}}_\theta(\alpha_{t'} \mathbf{x}_{\text{pr}} + \sigma_{t'} \boldsymbol{\epsilon}', \mathbf{c}_{\text{pr}}) - \mathbf{x}_{\text{pr}} \|_2^2 + L_{reg}]. \tag{8}$$

While training, we update the *concept neurons* and special prompt token following [18] by minimizing Eq. (8), without the need to store any additional model weights during continual learning. For inference, CNS can be directly applied to produce images with concepts observed anytime during training, users do not spend any additional computational effort such as test-time optimization.

5. Experiment

5.1. Experimental setup

Dataset. We conduct experiments on a dataset containing 20 concepts, composed with 8 real-world animals, 6 real-world objects, 3 styles and 3 real-world scenes. More details are provided in the supplementary.

The authors from NTU downloaded, evaluated, and completed the experiments on the datasets.

Figure 4. **Qualitative visualization.** Note that only Continual Diffusion [8] and CNS are capable of performing continual personalization, while Mix-of-Show [13] and Orthogonal Adaptation [27] require to keep LoRAs for each concept for personalization. It can be seen that our personalized outputs match concepts learned across different time, alleviating appearance leakage and catastrophic forgetting problems.

Implementation details. We leverage Stable Diffusion (SD-1.5[1]) as our pretrained text-to-image generation model to conduct comparison experiments. We fine-tune both the text embeddings and *concept neurons* using the Adam optimizer, with a learning rate of $5e-4$ for text embeddings and $3e-5$ for *concept neurons*. The training steps are 500 for a single concept, which takes about four minutes on a single 3090 GPU. For all the other methods mentioned in this paper, we use the same backbone and pretrained weight as ours. For more implementation details, we provide them in the supplementary.

Evaluation metrics. Following [18], we use CLIP [28] to evaluate image- and text-alignment. For single-concept image alignment, both the generated and concept images are fed into the CLIP image encoder to obtain embeddings, and their cosine similarity is calculated. For multi-concept

[1] https://huggingface.co/runwayml/stable-diffusion-v1-5

Table 1. **Quantitative comparisons of single and multi-concept personalization.** In addition to the alignment-based metrics of CLIP-I and CLIP-T, we provide the computation estimates for different personalization methods. Note that memory requirements for GPU/CPU and computation time for multi-concept personalization indicate the *additional* costs for fusing concept weights previously learned.

Methods	Single Concept		Multiple Concepts		Computational Resources	
	CLIP-I ↑	CLIP-T↑	CLIP-I↑	CLIP-T↑	Memory(MB)↓	Time(s)↓
Textual Inversion [11]	72.76	72.69	65.30	65.00	**0 / 0**	**0**
Custom Diffusion [18]	67.88	74.92	65.87	68.70	3547 / **0**	10
Mix-of-Show [13]	**75.86**	75.75	65.26	70.62	62852 / **0**	727
Orthogonal Adaption [27]	74.67	74.87	66.37	69.20	5663 / 3167	42
Continual Diffusion [32]	71.82	66.12	66.15	60.30	2461 / 4747	10
CNS	74.88	**76.95**	**67.21**	**79.22**	**0 / 0**	**0**

Table 2. **Ablation study of our approach.** We compare CNS with two baselines, removing the continual regularization loss in Sec. 4.3 and randomly picking fine-tuned neurons (i.e., no *concept neurons* selection of Sec. 4.2).

Components			Single Concept	
random neurons	concept neurons	L_{reg}	CLIP-I↑	CLIP-T↑
✓		✓	73.15	73.30
	✓		73.06	74.65
	✓	✓	**74.88**	**76.95**
			Multiple Concepts	
✓		✓	65.05	72.92
	✓		66.89	77.02
	✓	✓	**67.21**	**79.22**

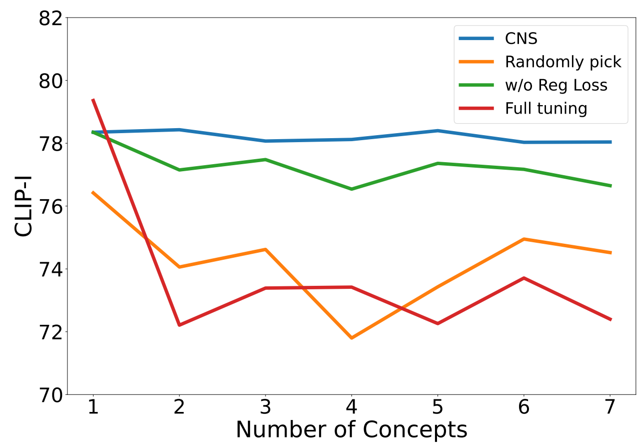

Figure 5. **Performance degradation on the first concept over time.** Compared to the ablated version of our method, the full version of CNS is sufficiently robust during continual learning, resulting in negligible degradation on the first concept. This confirms our ability in alleviating catastrophe forgetting problems.

personalization, image alignment is measured as the average visual similarity between the generated image and each concept image. To assess text similarity, the CLIP image encoder processes the generated images and the CLIP text encoder processes the text prompt. The cosine similarity between these embeddings serves as the text alignment score.

5.2. Qualitative comparisons

We demonstrate the result of CNS and other competitive methods in Fig. 4. For Mix-of-show [13] and Orthogonal Adaptation [27], we learned LoRA for each concept and merged them by their proposed fusion methods. For the others, we incrementally learn from the first concept to the seventh concept. As shown in Fig. 4, CNS outperforms all the other methods in both single-concept and multi-concept personalization. For single-concept setting, CNS is able to match the prompt perfectly and consistently. As more concepts are included in prompts, the other three methods suffer from concept vanish or attribute binding (*e.g.* Continual Diffusion generates a <new1> dog-like horse even though <new1> dog is not mentioned in the prompt.). Yet, images generated by CNS still preserve the image-alignment for all concepts and text-alignment for input prompt.

5.3. Quantitative comparisons

Quantitative comparisons of image quality. We evaluate 3 continual personalization sets, each set contains 7 concepts including animals, styles, objects and scenes. Each set contains 35 single-concept prompts, 7 multi-concept prompts. We generate 50 images per prompt and pick the best as representative image.

As shown in the first column of Tab. 1, CNS is slightly lower than Mix-of-Show [13] on single-concept image-alignment yet attains the highest score on single-concept text-alignment, which indicates that CNS does not suffer from overfitting the concept image, and it effectively captures semantic details from text descriptions. CNS outperforms all the other method in multi-concept image-alignment and text-alignment. It is worth mentioning that all the other methods suffer from performance degradation on multi-concept text-alignment compared to single-concept. In the contrast, Table 1 shows that CNS preserves text-to-image ability to multi-concept generation and achieve the highest image quality.

Quantitative comparisons of computational efficiency. Table 1 compares memory and time consumption across various fusion methods when fusing seven concepts. Notably, while Mix-of-Show [13] achieves near-top scores in both CLIP-I and CLIP-T, it also demands the highest memory and time resources, highlighting a trade-off between fusion efficiency and performance in prior methods. In contrast, CNS demonstrates balanced fusion efficiency and high performance, eliminating the need for such compromises. Additionally, to avoid concept fusion issues, methods like [13, 27] combine LoRA weights for each concept within a single image, though this requires multiple fusion operations for different concept combinations.

5.4. Ablation Study

Quantitative ablation study. In Tab. 2, we examine the effects of regularization loss and targeted fine-tuning of concept neurons. To assess the impact of fine-tuning *concept neurons*, we randomly selected the same number of neurons as *concept neurons* and fine-tuned them. As shown in the first row of Tab. 2, both CLIP-I and CLIP-T scores decrease in single- and multi-concept scenarios, indicating that our neuron selection method successfully identifies the most representative neurons, enhancing personalization performance. Additionally, we ablate the regularization loss L_{reg}, which prevents catastrophic forgetting by preserving overlapping concept neurons with previous concept and pretrained weights. Without L_{reg}, some overlapping neurons are overwritten by new concepts, leading to catastrophic forgetting and degraded performance.

Catastrophic forgetting ablation study. To prove that our CNS framework is robust enough for preventing from catastrophic forgetting, we conduct an experiment as depicted in Fig. 5. We illustrate the degradation of CLIP-I score of the first learned concept after learning multiple concepts incrementally. In this experiment, we report the degradation curves of the four baselines: the green curve represents ours method without continual regularization loss as detailed in Sec. 4.3; the yellow curve represents that we fine-tune same amount of randomly picked neurons as our selected *concept neurons* as detailed in Sec. 4.2. The red curve represents fine-tune all the parameters of key and value mapping from the text latent in the cross-attention layers. The results show that with *concept neurons* selection and the continual regularization loss, CNS can realize continual personalization while preventing from catastrophe forgetting effectively comparing to all the baselines.

Table 3. **Human study.** Scores for image and text alignments denote the percentages of users who considered the image generated by the corresponding method to be the most desirable.

Methods	Image Alignment(%)	Text Alignment(%)
Textual Inversion [11]	3.1	4.7
Custom Diffusion [18]	1.2	0.8
Mix-of-Show [13]	20.2	22.3
Orthogonal Adaption [27]	14.3	16.2
Continual Diffusion [32]	1.4	0.9
CNS	**59.8**	**55.1**

Human evaluation. We hire 10 users to conduct human study to further evaluate CNS. In the experiment, we provide 5 single-concept and 15 multi-concept prompts, and generate 5 samples for each prompt. For each question, users are required to select the image aligned mostly with the input prompt and the image most similar to the target concepts' images across all samples generated by each method. The results are shown in Tab. 3.

Further analysis. We present additional results in the supplementary. For example, we conduct experiments to confirm our scoring mechanism, and those verifying the selected neurons during continual personalization. And, experiments with region control show that CNS can be integrated with different text-to-image generation models.

6. Conclusion and Future Work

We present Concept Neuron Selection, CNS, a simple yet effective approach for continual learning-based personalization. By identifying concept-specific neurons in LDMs, CNS incrementally fine-tunes these neurons while preserving prior knowledge and zero-shot text-to-image generation. CNS achieves state-of-the-art performance with minimal parameter updates, outperforming existing methods in single- and multi-concept scenarios. Additionally, CNS operates fusion-free, reducing memory and processing demands for continual personalization.

Compared to previous continual personalization approaches which require to store LoRAs for each concept, we only need to store the concept neuron masks $\mathbf{M}_{1:m}$ to track previously tuned neurons, with significantly less memory requirement. As future research directions, our proposed concept neuron selection scheme can be possibly extended to tackle knowledge editing and unlearning tasks, not limited to particular data modality.

7. Acknowledgment

This work is supported in part by the National Science and Technology Council via grant NSTC113-2634-F-002-005, and in part by Qualcomm Technologies, Inc. through a Taiwan University Research Collaboration Project.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 4

[2] Omer Antverg and Yonatan Belinkov. On the pitfalls of analyzing individual neurons in language models. In *ICLR*, 2022. 3

[3] Ruchika Chavhan, Da Li, and Timothy Hospedales. Conceptprune: Concept editing in diffusion models via skilled neuron pruning. *arXiv preprint arXiv:2405.19237*, 2024. 3, 4

[4] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. In *SIGGRAPH*, 2023. 1

[5] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. arxiv abs/2307.09481 (2023). In *CVPR*, 2024. 2

[6] Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, Anthony Bau, and James Glass. What is one grain of sand in the desert? analyzing individual neurons in deep nlp models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019. 3

[7] Ganggui Ding, Canyu Zhao, Wen Wang, Zhen Yang, Zide Liu, Hao Chen, and Chunhua Shen. Freecustom: Tuning-free customized image generation for multi-concept composition. In *CVPR*, 2024. 2

[8] Jiahua Dong, Wenqi Liang, Hongliu Li, Duzhen Zhang, Meng Cao, Henghui Ding, Salman Khan, and Fahad Shahbaz Khan. How to continually adapt text-to-image diffusion models for flexible customization? In *NIPS*, 2024. 2, 3, 6, 1

[9] Nadir Durrani, Hassan Sajjad, Fahim Dalvi, and Yonatan Belinkov. Analyzing individual neurons in pre-trained language models. In *EMNLP*, 2020. 3

[10] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *ICLR*, 2023. 1

[11] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023. 1, 2, 7, 8

[12] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Encoder-based domain tuning for fast personalization of text-to-image models. In *SIGGRAPH*, 2023. 2

[13] Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, et al. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. In *NeurIPS*, 2024. 2, 6, 7, 8, 1, 4

[14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022. 2

[15] Sangwon Jang, Jaehyeong Jo, Kimin Lee, and Sung Ju Hwang. Identity decoupling for multi-subject personalization of text-to-image models. In *NIPS*, 2024. 1

[16] Jimyeong Kim, Jungwon Park, and Wonjong Rhee. Selectively informative description can reduce undesired embedding entanglements in text-to-image personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2

[17] Zhe Kong, Yong Zhang, Tianyu Yang, Tao Wang, Kaihao Zhang, Bizhu Wu, Guanying Chen, Wei Liu, and Wenhan Luo. Omg: Occlusion-friendly personalized multi-concept generation in diffusion models. In *ECCV*, 2024. 2

[18] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1, 2, 4, 5, 6, 7, 8

[19] Gihyun Kwon, Simon Jenni, Dingzeyu Li, Joon-Young Lee, Jong Chul Ye, and Fabian Caba Heilbron. Concept weaver: Enabling multi-concept fusion in text-to-image models. In *CVPR*, 2024. 2

[20] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023. 1

[21] Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *NIPS*, 2023. 2

[22] Zhiheng Liu, Ruili Feng, Kai Zhu, Yifei Zhang, Kecheng Zheng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones: Concept neurons in diffusion models for customized generation. In *ICML*, 2023. 3

[23] Zhiheng Liu, Yifei Zhang, Yujun Shen, Kecheng Zheng, Kai Zhu, Ruili Feng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones 2: Customizable image synthesis with multiple subjects. In *NeurIPS*, 2023. 3

[24] Wan-Duo Kurt Ma, Avisek Lahiri, John P Lewis, Thomas Leung, and W Bastiaan Kleijn. Directed diffusion: Direct control of object placement through attention guidance. In *AAAI*, 2024. 1

[25] Tuna Han Salih Meral, Enis Simsar, Federico Tombari, and Pinar Yanardag. Clora: A contrastive approach to compose multiple lora models. *arXiv preprint arXiv:2403.19776*, 2024. 2

[26] Saman Motamed, Danda Pani Paudel, and Luc Van Gool. Lego: Learning to disentangle and invert concepts beyond object appearance in text-to-image diffusion models. *arXiv e-prints*, 2023. 2

[27] Ryan Po, Guandao Yang, Kfir Aberman, and Gordon Wetzstein. Orthogonal adaptation for modular customization of diffusion models. In *CVPR*, 2024. 2, 6, 7, 8, 1

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 6

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1

[30] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 1, 2, 5

[31] Viraj Shah, Nataniel Ruiz, Forrester Cole, Erika Lu, Svetlana Lazebnik, Yuanzhen Li, and Varun Jampani. Ziplora: Any subject in any style by effectively merging loras. In *European Conference on Computer Vision*, 2025. 2

[32] James Seale Smith, Yen-Chang Hsu, Lingyu Zhang, Ting Hua, Zsolt Kira, Yilin Shen, and Hongxia Jin. Continual diffusion: Continual customization of text-to-image diffusion with c-lora. In *TMLR*, 2023. 2, 3, 7, 8

[33] Xavier Suau, Luca Zappella, and Nicholas Apostoloff. Finding experts in transformer models. *arXiv preprint arXiv:2005.07647*, 2020. 3

[34] Gan Sun, Wenqi Liang, Jiahua Dong, Jun Li, Zhengming Ding, and Yang Cong. Create your world: Lifelong text-to-image diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 3

[35] Mingjie Sun, Zhuang Liu, Anna Bair, and J Zico Kolter. A simple and effective pruning approach for large language models. In *ICLR*, 2024. 4

[36] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 3

[37] Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. p+: Extended textual conditioning in text-to-image generation. *arXiv preprint arXiv:2303.09522*, 2023. 2

[38] Xiaozhi Wang, Kaiyue Wen, Zhengyan Zhang, Lei Hou, Zhiyuan Liu, and Juanzi Li. Finding skill neurons in pre-trained transformer-based language models. 2022. 3

[39] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *ICCV*, 2023. 1

[40] Xun Wu, Shaohan Huang, and Furu Wei. Mixture of lora experts. In *ICLR*, 2024. 2

[41] Yang Yang, Wen Wang, Liang Peng, Chaotian Song, Yao Chen, Hengjia Li, Xiaolong Yang, Qinglin Lu, Deng Cai, Boxi Wu, et al. Lora-composer: Leveraging low-rank adaptation for multi-concept customization in training-free diffusion models. *arXiv preprint arXiv:2403.11627*, 2024. 2

[42] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. In *TMLR*, 2023. 1

[43] Yanbing Zhang, Mengping Yang, Qin Zhou, and Zhe Wang. Attention calibration for disentangled text-to-image personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2

[44] Zhengyan Zhang, Yankai Lin, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. Moefication: Transformer feed-forward layers are mixtures of experts. In *ACL*, 2022. 3

[45] Zhengyan Zhang, Zhiyuan Zeng, Yankai Lin, Chaojun Xiao, Xiaozhi Wang, Xu Han, Zhiyuan Liu, Ruobing Xie, Maosong Sun, and Jie Zhou. Emergent modularity in pre-trained transformers. In *ACL*, 2023. 3