Adversarial Perturbation Shield: Preventing Concept Bleed-through in Continual Learning of Personalized Generative Models

Ziwen Lan[1], Keisuke Maeda[2,3], Takahiro Ogawa[2], Miki Haseyama[2]

[1]Graduate School of Information Science and Technology, Hokkaido University, Japan
[2]Faculty of Information Science and Technology, Hokkaido University, Japan
[3]Graduate School of Information Sciences, Tohoku University, Japan
lan@lmd.ist.hokudai.ac.jp, maeda@lmd.ist.hokudai.ac.jp, ogawa@lmd.ist.hokudai.ac.jp, mhaseyama@lmd.ist.hokudai.ac.jp

Abstract

Personalized text-to-image diffusion models have gained increasing attention because they can generate images that contain unique concepts based on limited training data. However, in continual learning scenarios, these models suffer from *concept bleed-through*, where newly introduced concepts frequently overwrite or interfere with the previously learned concepts. Previous studies have attempted to mitigate this issue at the model adaptation level; however, they failed to fully preserve the distinct semantic representations in the latent space. Thus, this paper proposes an adversarial perturbation-based training strategy to address concept bleed-through in continual learning for personalized diffusion models. The proposed method introduces adversarial perturbations into the training images, which strategically shifts their semantic representations in the latent space to ensure that the newly learned concepts remain distinct and do not interfere with the previously acquired knowledge. Unlike structural modifications to the model, the proposed method operates at the data level, which makes it broadly applicable to existing continual personalization frameworks without increasing model complexity. Experimental results demonstrate that the proposed method significantly improves concept separation while maintaining high image fidelity, offering a solution to enhance the reliability of continual learning in personalized generative models.

Introduction

Recent studies have investigated the use of diffusion models (Sohl-Dickstein et al. 2015) to generate high-quality images. Text-to-image diffusion models, e.g., the DALL-E 2 (Ramesh et al. 2022) and stable diffusion (Rombach et al. 2022a) models, have become the mainstream in research on text-to-image models, and these models are currently used for various tasks, including image generation from arbitrary text, image editing (Kawar et al. 2023; Zhang and Agrawala 2023), super-resolution (Gao et al. 2023), and inpainting (Lugmayr et al. 2022; Brooks, Holynski, and Efros 2023). Despite the extensive training of diffusion models, they may struggle to generate unique or personalized concepts that are not present in the training corpus, e.g., personalized styles or specific faces. There has been an increasing trend toward implementing personalization methods in text-to-image diffusion models, including textural inversion (Gal et al. 2022), DreamBooth (Ruiz et al. 2022), and LoRA (Hu et al. 2022). In addition to these methods, which are employed to realize personalization to a single concept, custom diffusion methods (Kumari et al. 2023) learn multiple concepts. These methods enable the creation of images related to a new specific concept (an object or style) based on a few reference examples, and the learning process is efficient by changing only a small portion of the weights in the entire pipeline of the diffusion model, which results in both fast concept acquisition and lightweight model updates.

These methods have demonstrated impressive results in generating personalized content; however, they suffer from a critical limitation when applied to the continual learning context, which is a common use case for personalized generative models. In continual learning, the models must learn multiple concepts sequentially over time without previous training data. This paradigm is essential to reduce computational cost and preserve data privacy because storing and retraining on previous data may be infeasible. However, existing methods for personalized diffusion models struggle with catastrophic forgetting, where newly learned concepts overwrite previously acquired ones, and catastrophic-neglect, where images generated by the models miss key concepts mentioned in the prompt. To address these issues, recent studies have introduced methods, e.g., C-LoRA (Smith et al. 2023a) and CIDM (Dong et al. 2024), which enhance the continual learning for diffusion models. These methods have demonstrated promising results; however, they still face a fundamental problem, which we call concept bleed-through (Smith et al. 2023a). As shown in Fig. 1, concept bleed-through occurs when learning a new concept inadvertently alters or distorts the generation of similar previously learned concepts, thereby leading to unintended hybrid or incorrect outputs.

There have been attempts to address concept bleed-through at the model structure level.However, an effective solution to prevent concept bleed-through while maintaining high-quality personalized image generation remains an open problem. In addition, these structural modifications are not directly applicable to existing personalization frameworks and frequently introduce additional model complexity, which limits their practicality in real-world applications.

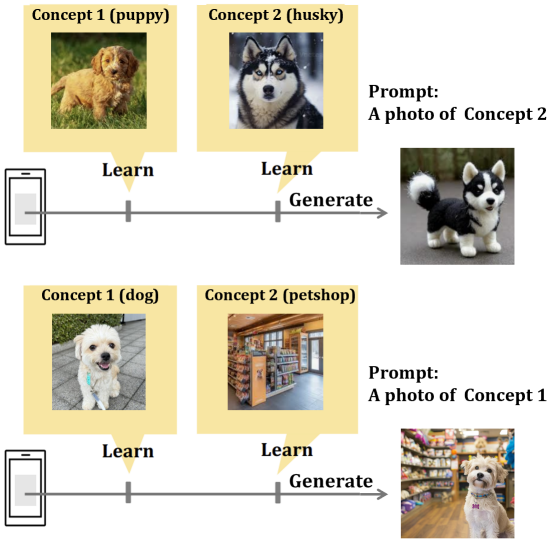

Figure 1: Examples of the concept bleed-through problem in the continual learning of personalized generative models. When handling nearby concepts (e.g., *puppy* and *husky*) or related concepts (e.g., *dog* and *petshop*), although the prompt only involves a single learned concept, the generated image is clearly influenced by the other learned concepts.

Thus, we focus on preprocessing the training images.

In the continual learning of diffusion models, handling similar concepts is challenging because the semantic features contained in images are inherently similar, which makes it difficult to avoid mapping them to closely positioned features in the latent space. Thus, we consider methods to manipulate the semantic information of images while preserving their visual information as much as possible. Inspired by recent advances in adversarial attacks on diffusion models (Shan et al. 2024), we propose a method to mitigate the concept bleed-through problem in personalized diffusion models. Adversarial attacks on diffusion models typically involve the introduction of carefully crafted perturbations to the input images to alter their semantic representations in a manner that confounds the model while remaining imperceptible to human observers. These adversarial attacks exploit the vulnerabilities of the diffusion models by modifying the latent representations of the training data, thereby leading to significant deviations in the generated outputs.

In this paper, we propose a training method for personalized diffusion models using adversarial perturbation. Specifically, we inject subtle adversarial perturbations into the training images used to personalize the diffusion model. These perturbations are designed to shift the semantic representation of the concepts in the latent space, which ensures that newly learned concepts occupy distinct regions that do not interfere with the previously learned concepts. By adopting this strategy, the proposed method effectively prevents concept bleed-through while preserving the fidelity and consistency of personalized image generation.

The technical contributions of this study are summarized as follows.

- We introduce a training strategy based on adversarial perturbation for personalized diffusion models, where adversarial perturbations are leveraged to prevent concept bleed-through in continual learning scenarios. To the best of our knowledge, this is the first study to introduce adversarial perturbations into the continual learning task of diffusion models.
- The proposed method systematically shifts the semantic representation of personalized concepts within the latent space, which ensures that the newly learned concepts do not interfere with previously acquired concepts, thereby enhancing the ability of the diffusion models to learn new concepts over time without catastrophic forgetting or unintended concept mixing.
- The proposed method operates at the data level rather than modifying the structure of the model, which ensures broad compatibility with existing personalization frameworks while avoiding additional model complexity.

Related Works

Personalization in Text-to-Image Diffusion Models

Text-to-image generation(Zhang et al. 2023), which requires training on a large corpus of text and an image paired dataset, has become increasingly popular due to the advancement in diffusion models (Ho, Jain, and Abbeel 2020; Rombach et al. 2022b; Croitoru et al. 2023). The trained models excel at producing diverse and realistic images according to user-specific input text prompts. However, generally, trained text-to-image models cannot handle novel personalized concepts, e.g., an individual's personal item or a particular individual's face. Text-to-image personalization aims to guide a diffusion-based text-to-image model to generate novel user-provided concepts through free text. In this process, the user provides image examples of a concept, which are then used to generate novel images containing the newly acquired concepts through text prompts. For example, DreamBooth (Ruiz et al. 2022) fine-tunes the entire parameter sets in a diffusion model using the given images of the new concept, and the textual inversion method (Gal et al. 2022) only learns the custom feature embedding "words." Note that these methods are employed to realize customization to only a single concept. In contrast, the custom diffusion (Kumari et al. 2023) and ED-LoRA (Gu et al. 2023) methods learn multiple concepts using a combination of cross-attention fine-tuning, regularization, and closed-form weight merging. However, these methods struggle to learn concepts in a sequential manner, motivating the need for the continual learning of personalized diffusion models.

Continual Learning

In continual learning, a model is trained on a sequence of tasks without forgetting previously learned knowledge, where each task has a different data distribution. Generally, existing methods can be classified into three categories based on their approach to mitigating catastrophic forgetting (McCloskey and Cohen 1989), i.e., regularization-based

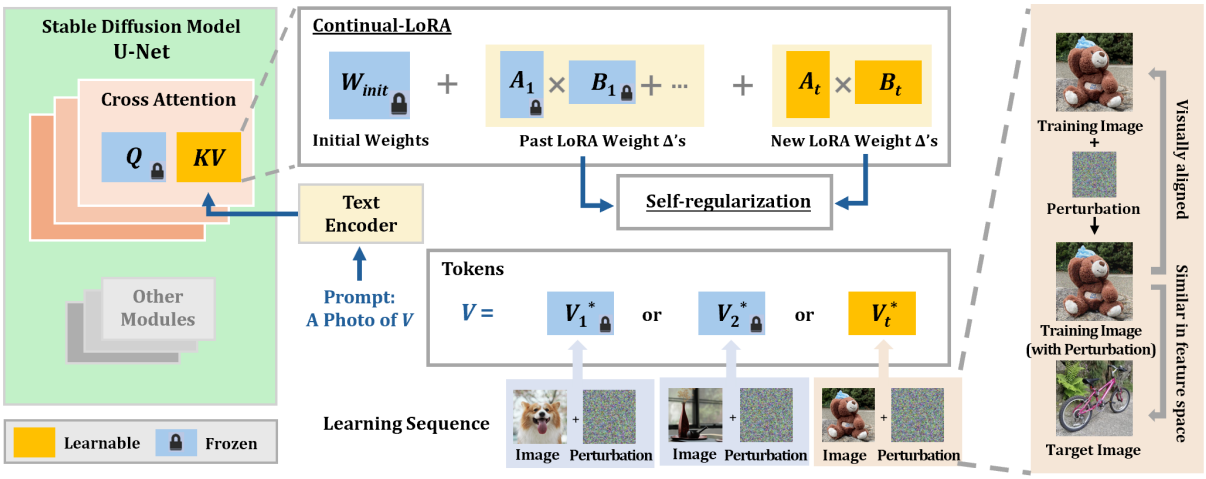

Figure 2: Overview of the proposed method. We update the key-value (K-V) projection in the U-Net cross-attention modules of stable diffusion using a continual, self-regulating low-rank weight adaptation. In addition, we add adversarial perturbations to the training data to achieve effective concept separation in the latent space.

methods (Douillard et al. 2020; Kirkpatrick et al. 2017; Li and Hoiem 2017), rehearsal-based methods (Pham, Liu, and Hoi 2021; Smith et al. 2021; Van de Ven, Siegelmann, and Tolias 2020), and prompt-based methods (Wang et al. 2022; Wang, Huang, and Hong 2022; Smith et al. 2023b). Regularization-based methods (Douillard et al. 2020; Kirkpatrick et al. 2017; Li and Hoiem 2017) add extra regularization terms to the objective function while learning a new task. For example, the elastic weight consolidation (Kirkpatrick et al. 2017) method estimates the importance of the parameters and applies a per-parameter weight decay. In addition, rehearsal-based methods (Pham, Liu, and Hoi 2021; Smith et al. 2021; Van de Ven, Siegelmann, and Tolias 2020) store or generate samples from previous tasks in a data buffer and replay them with new task data; however, this may not always be feasible due to privacy or copyright concerns. Prompt-based continual learning methods for vision Transformers, e.g., L2P (Wang et al. 2022), S-Prompt (Wang, Huang, and Hong 2022), and CODA-Prompt (Smith et al. 2023b), have been shown to outperform rehearsal-based methods without requiring a replay buffer. These methods are used for classification problems, and recently generative prompt-based continual learning methods, e.g., Lifelong-GAN (Zhai et al. 2019) and C-LoRA (Smith et al. 2023a), have been proposed. In the current study, we focus on these prompt-based methods.

Adversarial Attacks and Adversarial Perturbations

With the introduction of the fast gradient sign method attack (Goodfellow, Shlens, and Szegedy 2014), adversarial vulnerability has become an active field of research in the machine learning domain. The goal of adversarial attacks is to generate a model input (typically by adding carefully designed tiny perturbations into the data) that can induce a misclassification while remaining visually indistinguishable from a clean one. Following this foundational work, different attacks that utilize various approaches have emerged. Unlike the adversarial attack for the classification models, in this study, we focus on generative models. Although adversarial example generation is a diverse and well-studied field in classification, projected gradient descent (PGD) (Madry et al. 2018) remains the dominant approach in image generation, and to the best of our knowledge, no well-established alternatives currently exist for generative diffusion models. For example, PGD-based methods Glaze (Shan et al. 2023) and Nightshade (Shan et al. 2024), which are attack methods for generative diffusion models, disrupt the generation process by changing the semantic information contained in the images.

Methodology

In this section, we introduce the proposed method in detail. An overview of the proposed method is shown in Fig. 2. It is worth clarifying that the first two subsections serve to explain the baseline continual learning framework in a simple diffusion model. The key novelty lies in last subection, where we introduce our adversarial perturbation strategy, and our aim is to isolate the contribution of adversarial perturbation under a simple continual learning framework.

Parameter Modification in Continual LoRA

Inspired by (Smith et al. 2023a), and considering that the current mainstream continuous learning frameworks for personalized image generation are mostly based on LoRA, we adopt a LoRA-based fine-tuning method and introduce adversarial perturbations to the training images to shift their semantic representations within the latent space. This ensures that the newly learned concepts do not interfere with the previous concepts, thereby preventing the concept bleed-through. We use LoRA due to its efficiency in updating only a small fraction of parameters while preserving most of the pretrained weights. Unlike full fine-tuning methods, which adjust all parameters, LoRA-based adaptation is considerably more memory-efficient and can be applied sequentially

without increasing the size of the model significantly. Thus, it is well-suited for continual learning scenarios where the model updates must be performed iteratively without catastrophic forgetting.

In the continual diffusion setting, we learn N personalization tasks $t \in \{1, 2, \ldots, N-1, N\}$, where N denotes the total number of concepts that will be learned by the model. Following the results of the custom diffusion process (Kumari et al. 2023), we modify a small subset of parameters in the stable diffusion model (Rombach et al. 2022a), specifically the cross-attention layers in U-Net because the cross-attention parameters of U-Net are most sensitive to changes during the personalization of the diffusion model. Here, we freeze the query features Q and parameterize the weight updates $W^{K,V} \in \mathbb{R}^{D_1 \times D_2}$ of the key features K and the value features V using LoRA (Hu et al. 2022):

$$\begin{aligned} W_t^{K,V} &= W_{t-1}^{K,V} + A_t^{K,V} B_t^{K,V} \\ &= W_{init}^{K,V} + \left[\sum_{t'=1}^{t-1} A_{t'}^{K,V} B_{t'}^{K,V} \right] + A_t^{K,V} B_t^{K,V}, \end{aligned} \tag{1}$$

where $A_t^{K,V} \in \mathbb{R}^{D_1 \times r}$ and $B_t^{K,V} \in \mathbb{R}^{r \times D_2}$ correspond to the key and value projection matrices at time step t, which encode the learned concept information. Note that r is a low-rank hyperparameter, and D_1 and D_2 denote the input and output dimensions of the key and value projection matrices in the cross-attention layers, respectively. Furthermore, $W_{init}^{K,V}$ is the inital values from the pre-trained model. The above approach allows for efficient parameter adaptation while maintaining previous knowledge.

Self-regularization

To mitigate catastrophic forgetting and preserve the previously learned concepts in personalized generative models, we introduce a self-regularization loss that constrains the parameter updates. This regularization term shown in the following equation ensures that updates affecting new concepts do not interfere destructively with the existing knowledge:

$$\mathcal{L}_{\text{forget}} = \| \left| \sum_{t'=1}^{t-1} A_{t'}^{K,V} B_{t'}^{K,V} \right| \odot A_t^{K,V} B_t^{K,V} \|_F^2, \tag{2}$$

where $\odot$ represents element-wise multiplication, and $\| \cdot \|_F^2$ denotes the squared Frobenius norm. This self-regularization loss penalizes overlapping activations between the previously learned (t') and the newly introduced (t) concepts in the latent space, which effectively reduces semantic interference.

By integrating the self-regularization into the training objective, the proposed method achieves an effective balance between continual adaptation and knowledge preservation, which enables the personalized diffusion model to learn a sequence of fine-grained concepts, thereby resulting in more stable image generation. Specifically, when learning the personalization task t, we minimize the following total loss:

$$\min_{W_t^{K,V} \in \theta} \mathcal{L}_{\text{diff}}(i, \theta) + \lambda \mathcal{L}_{\text{forget}}, \tag{3}$$

where i is the input training data of the new concept, $\mathcal{L}_{\text{diff}}$ is the loss function for the stable diffusion model θ (Rombach et al. 2022a), and λ is a hyperparameter selected with a simple exponential sweep.

Training on Images with Adversarial Perturbations

The self-regularization enhances the model's ability to resist catastrophic forgetting, ensuring that previously learned concepts in the learning sequence are retained; however, it does not effectively mitigate the adverse interactions between different concepts, particularly in cases where semantically similar concepts are introduced sequentially. As mentioned in the introduction, existing continual learning methods for generative models, e.g., C-LoRA and STAMINA, primarily focus on mitigating forgetting but do not explicitly address concept interference. To overcome this limitation, we introduce adversarial perturbations to the training images, which strategically shift the semantic features in the latent space while preserving the quality and visual consistency of the image. From a theoretical perspective, adversarial perturbations are designed to maximize the divergence between representations in the latent space by targeting directions of high sensitivity in the model's gradient. Specifically, the optimization encourages the perturbed inputs to move away from the existing concept's latent features, thereby enforcing semantic separation. Mathematically, this aligns with maximizing the cosine distance, akin to contrastive representation learning.

Inspired by previous studies (Shan et al. 2024), for an image x_t, we optimize the perturbation δ using targeted PGD (Madry et al. 2018):

$$\begin{aligned} \delta :&= \arg\min_{\|\delta\| \leqslant \epsilon} \mathcal{L}_{\mathcal{E}}(x_t, \delta, x_t^{\text{pr}}) \\ &= \arg\min_{\|\delta\| \leqslant \epsilon} \left\| \mathcal{E}(x_t^{\text{pr}}) - \mathcal{E}(x_t + \delta) \right\|_2^2, \end{aligned} \tag{4}$$

where ϵ is the perturbation budget, $\mathcal{E}$ and x_t^{pr} are the encoder of the diffusion model and the given target image for x_t, respectively. Note that the selection of x_t^{pr} is critical in terms of ensuring effective perturbation-induced separation in the latent space. Specifically, for each input image x_t associated with a new concept, we sample x_t^{pr} from a pool of images that are semantically dissimilar to $x_1, x_2, ..., x_{t-1}$, based on their precomputed latent representations. This ensures that the perturbation δ shifts the latent encoding of x_t away from the region associated with $x_1, x_2, ..., x_{t-1}$. As a result, we promote better separability between the learned concepts by encouraging the features of different concepts to occupy distinct and distant regions of the latent space.

By minimizing the loss $\mathcal{L}_{\mathcal{E}}$ in Eq. (4), the perturbed image $x_t + \delta$ is embedded into the latent space aligned with the target image, which effectively enforces a shift in the semantic representation. This adjustment prevents undesired overlaps between the newly learned concepts and the previously stored knowledge. By constraining the perturbation budget ϵ during optimization, the perturbation is kept below a certain threshold, preventing degradation in image quality or unintended changes in image semantics. Consequently, the proposed method significantly reduces the risk of concept

interference, which ensures that the personalized concepts remain distinct over time. The proposed method is novel because we explicitly manipulate the latent space via adversarial perturbations, thereby creating a more robust framework for continual learning in generative models.

To further mitigate the interference between concepts, we follow a previously proposed custom tokenization strategy (Smith et al. 2023a, 2024). Here, we define N personalized tokens $V_1^*, V_2^*, \ldots, V_N^*$, which are initialized randomly rather than using less frequent words. In addition, to avoid interference with existing concepts, the object names are removed from the input sequence. During inference, the learned embeddings replace these tokens to generate images of multiple learned concepts simultaneously. This tokenization strategy ensures that different concepts retain unique representations in both the textual and latent spaces, which reduces the likelihood of semantic entanglement.

It is worth noting that our method essentially establishes a connection between the customized text embeddings and the slightly misaligned image features through targeted guidance using the target image. Therefore, it does not compromise the model's original ability to generate the concepts present in the target image.

Experiments

Main Experimental Settings

To evaluate the effectiveness of the proposed adversarial perturbation-based training strategy, we performed experiments on personalized diffusion models under a continual learning setting. Here, the goal was to assess whether the proposed method mitigates concept bleed-through while maintaining high-quality continual personalized image generation. The quantity and content of the experiments followed those of (Smith et al. 2023a) and (Dong et al. 2024).

Implementation Details: Following previous studies that utilized existing methods (Kumari et al. 2023; Smith et al. 2023a), we used stable diffusion v1.5 (Rombach et al. 2022a) as the backbone. In addition, we utilized the prompt "a photo of a $V*$," where $V*$ is a learnable custom token. For LoRA, we searched for the rank using a simple exponential sweep and found that a rank of 16 learned all concepts sufficiently. During the training process, the learning rate was 2×10^{-6}, and the number of training steps performed on each image was set to 100.

In the optimization of the perturbation, we fixed the l_∞-norm as the constraint to obtain all adversarial images. During the optimization process, we experimentally set the sampling step to 40, the total perturbation budget to 8/255, and the batch size to 4.

Datasets: To evaluate the effectiveness of the proposed method, we used the official DreamBooth dataset (Ruiz et al. 2022), which comprises 30 distinct concept classes, each containing 5–6 images. Additionally, we also adopted the CustomConcept101 dataset (Kumari et al. 2023), which comprises 101 distinct concept classes, each containing 3–15 images. These dataset are commonly used to train personalized diffusion models. Also, to optimize the adversarial perturbation, we employed target images sampled from ImageNet dataset (Deng et al. 2009). **Compared methods**: The proposed method was compared with the following personalization methods for generative diffusion models: 1) Textual inversion (Gal et al. 2022), DreamBooth (Ruiz et al. 2022), InstantBooth (Shi et al. 2024): baseline personalization methods for single concept; 2) Custom Diffusion (Kumari et al. 2023), ED-LoRA (Gu et al. 2023): mechanisms for training models that can generate multiple concepts simultaneously; 3) C-LoRA (Smith et al. 2023a), STAMINA (Smith et al. 2024), CIDM (Dong et al. 2024): SOTA of continual learning methods for personalized diffusion models.

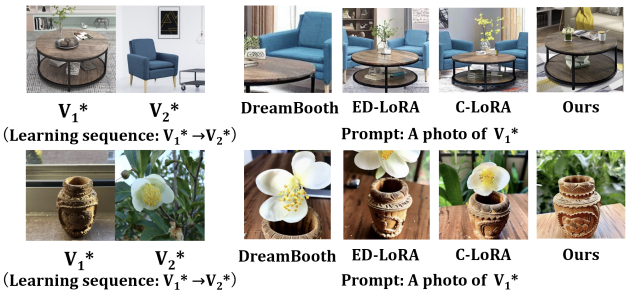

Figure 3: Qualitative results of experiments. All compared methods were affected by concept bleed-through in the learning sequence when dealing with both the nearby concepts (e.g., table and chair) and the related concepts (e.g., vase and flower). In contrast, our proposed method ensured accurate representation of the concepts in the image.

Method	$A_{mmd}(\downarrow)$	$F_{mmd}(\downarrow)$	CLIP Score(↑)
Textual inversion	2.98	**0.00**	0.71
DreamBooth	9.01	8.25	0.75
InstantBooth	3.34	6.91	0.74
Custom diffusion	5.57	3.96	0.78
ED-LoRA	7.72	7.08	0.82
C-LoRA	2.95	0.23	0.80
STAMINA	2.34	0.27	0.83
CIDM	2.11	0.22	0.85
C-LoRA + Ours	1.99	0.32	0.84
STAMINA + Ours	**1.92**	0.26	0.85
CIDM + Ours	1.98	0.22	**0.87**

Table 1: Comparison of MMD and CLIP scores across different methods.

Evaluation metrics: To assess the effectiveness of the proposed method, we employed evaluation metrics that focus on the quality of the generated images and the effectiveness of the perturbations in terms of preventing concept bleed-through. Specifically, we used Maximum mean discrepancy (MMD) score (Gretton et al. 2012) and CLIP score (Radford et al. 2021) to evaluate the performance of the methods. Here, we defined A_{mmd} as the average MMD score ($\times 10^3$) across all concepts after completing the entire training sequence, and F_{mmd} measured the average increase of MMD score for the previously learned concepts after learning subsequent concepts, quantifying the degree of forgetting.

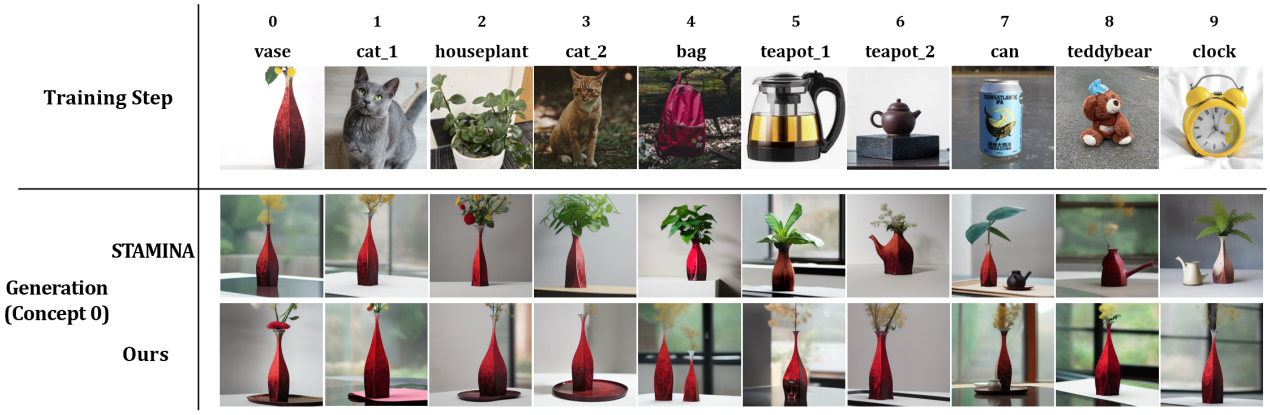

Figure 4: Examples of the additional experiment with different numbers of concepts in the training images. We trained the model with ten concepts in the learning sequence. We selected STAMINA (Smith et al. 2024) as the baseline method, and after each new concept was learned, the model was used to generate images of the most initially learned concept ($vase$).

Experimental Results

Table 1 shows the quantitative results for each evaluated method. Note that all results are averaged over 100 different multiconcept learning patterns, where each pattern contains two sequentially learned personalization tasks. As can be seen, the proposed method achieved the highest CLIP score and the lowest A_{mmd}, demonstrating improved concept preservation and better alignment with textual prompts. In addition, in terms of the F_{mmd} results, the proposed method also performed well relative to preventing forgetting (note that textual inversion (Gal et al. 2022) has no forgetting because its backbone is completely frozen). These results suggest that adversarial perturbations can play a crucial role in enhancing the stability and reliability of diffusion-based continual learning models. It is also worth mentioning that our method exhibited significant results when applied to C-LoRA, STAMINA, and CIDM, proving that the proposed method is widely compatible with existing personalization frameworks. Figure 3 shows a qualitative comparison of the images generated using the different methods. As can be seen, the proposed method produced images that exhibit clearer concept boundaries and improved semantic alignment, particularly in challenging generation tasks.

Additional Experiments

In addition to the main experiments, we designed a series of additional experiments to further validate the effectiveness of the proposed method. Here, if not specifically stated, the experimental settings followed the main experiment. If not mentioned, we only applied our method to (Dong et al. 2024) in all additional experiments.

In the additional experiments, we manipulated the key variables as follows.

Number of Concepts in Training Images. A common challenge in continual learning for personalized image generation is the increasing risk of concept bleed-through as more concepts are introduced. To evaluate this, we followed the setup commonly seen in DreamBooth studies (Ruiz et al. 2022), where up to ten concepts are trained sequentially (in this experiment, the concept sequence was $vase$, cat_1, $houseplant$, cat_2, bag, $teapot\_1$, $teapot\_2$, can, $teddybear$, $clock$). The results are shown in Fig. 4. As can be seen, without appropriate mitigation strategies, concept interference becomes more pronounced as the number of concepts increases, leading to a degradation in the image generation quality. By integrating our adversarial perturbation strategy, we observed a significant improvement in terms of concept separation. Specifically, the images generated after learning all 10 concepts exhibited reduced semantic drift compared with the baseline methods without adversarial perturbations.

Categories of Target Images. The selection of the target image categories for adversarial perturbation optimization plays a crucial role in determining the effectiveness of the proposed method. Thus, we investigated different target image categories sampled from ImageNet, including images that do not overlap conceptually with the training data in the entire learning sequence and images that are semantically related to the training data. In this additional experiment, we employed learning sequence ($table \rightarrow chair$) that was utilized in the main experiment. Specifically, we fixed the learning sequence to $table \rightarrow chair$, and only inject perturbation to the training images for $chair$. Figure 5 shows the generation outcomes when training images for $chair$ are paired with target images from five different categories: $vase$, dog, $airplane$, $bread$, and $microwave$ in the perturbation optimization. We also measured the cosine similarity between the generated image and both the $table$ (V1) and $chair$ (V2) embeddings.

Qualitative results (Fig. 5) indicate that pairing $chair$ with highly dissimilar targets (e.g., dog, $bread$) yields clearer and more separable visual representations. Conversely, semantically closer pairs (e.g., $vase$, $microwave$) show subtle feature blending. These results align with the quantitative findings in Table 2. These results further validate the principle that semantic dissimilarity between training and target images promotes stronger adversarial-induced latent separation.

Target Image Category	Similarity with V1 (Table)↓	Similarity with V2 (Chair)↓	A_{mmd} ↓	F_{mmd} ↓	CLIP ↑
-	-	-	3.12	0.51	0.76
vase	0.814	0.701	2.82	0.41	0.81
dog	**0.187**	0.325	1.92	**0.25**	0.85
airplane	0.523	0.544	1.96	0.28	0.84
bread	0.236	**0.297**	**1.91**	0.26	**0.86**
microwave	0.692	0.610	1.94	0.30	0.84

Table 2: Performance of our method on different target images for the perturbation optimization of $chair$. "-" means we do not adopt any perturbation to the training images.

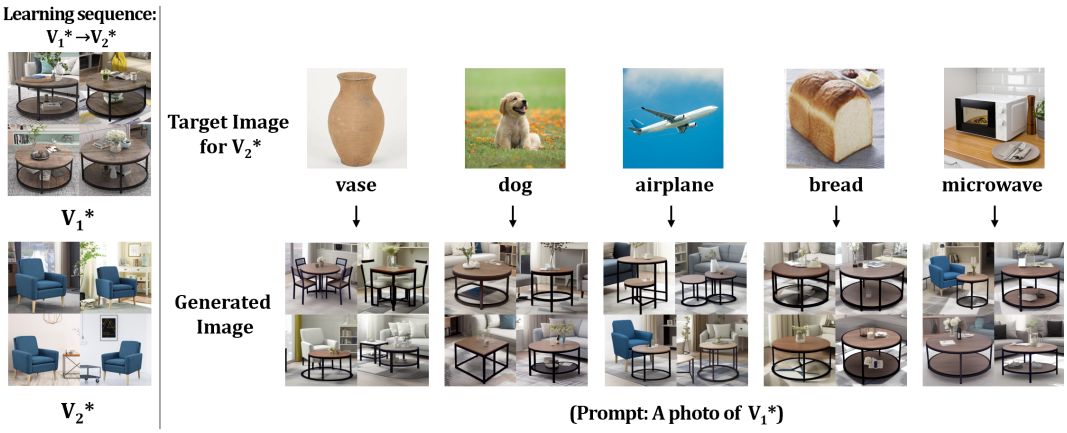

Figure 5: Examples of generated images using different target image categories for perturbation. Semantically dissimilar pairs lead to better concept separation.

Method	Chair CLIP Score ↑	Table CLIP Score ↑
C-LoRA	0.782	0.741
C-LoRA (w/SA)	0.851	0.094
C-LoRA (w/Ours)	0.847	0.822

Table 3: Comparation between selective forgetting methods and ours method.

Comparation with Selective Forgetting Methods

In this section, we discuss works related to concept erasure and selective forgetting, which are closely connected to the topic of concept bleed-through. It is worth noting that while these methods share a similar objective (mitigating interference between concepts), their application scenarios and underlying philosophies differ significantly from our proposed method. Selective forgetting methods are typically designed to intentionally erase previously learned concepts before introducing a new one. This strategy is effective for avoiding interference but comes at the cost of losing the model's ability to generate forgotten concepts. In contrast, our method does not require any forgetting steps and instead prevents concept bleed-through proactively during training. This allows the model to maintain the ability to generate all previously learned concepts, offering better flexibility and generalization in continual learning settings.

Despite these decisive differences, we still conducted additional comparative experiments against a representative forgetting-based method: Selective Amnesia (SA) (Heng and Soh 2023). Following the training sequence shown in Fig. 4 of our main manuscript ("$table \rightarrow chair$"), we compared the quality of generation of $chair$ after continual training. We use C-LoRA as the baseline method. In the case of Selective Amnesia, the model is explicitly fine-tuned to forget the concept of $table$ before learning $chair$, whereas our method integrates adversarial perturbations to preserve separation without forgetting. Table 3 shows the CLIP similarity scores between generated images and their corresponding text prompts.

The results indicate that both methods achieve comparable performance on the new concept ($chair$). However, our method is able to retain the model's ability to generate the previous concept ($table$). This demonstrates that our method achieves a similar level of interference mitigation without sacrificing the knowledge of learned concepts.

Conclusion

This paper has proposed an adversarial perturbation-based training strategy to mitigate the concept bleed-through issue in personalized generative models. By injecting a carefully designed adversarial perturbation into the training images, the proposed method shifts the semantic representations in the latent space, thereby ensuring a clear separation of the new and previously learned concepts. The proposed method was evaluated experimentally, and the results demonstrated the effectiveness of the method through contrastive learning and latent space alignment strategies, which help preserve the integrity of the learned concepts.

Acknowledgments

This work is supported in part by JSPS KAKENHI Grant Number JP23K21676, JP23K11211, and JST BOOST, Japan Grant Number JPMJBS2426.

References

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. InstructPix2Pix: Learning to Follow Image Editing Instructions. *arXiv preprint arXiv:2211.09800*.

Croitoru, F.-A.; Hondru, V.; Ionescu, R. T.; and Shah, M. 2023. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 248–255.

Dong, J.; Liang, W.; Li, H.; Zhang, D.; Cao, M.; Ding, H.; Khan, S. H.; and Shahbaz Khan, F. 2024. How to continually adapt text-to-image diffusion models for flexible customization? *Advances in Neural Information Processing Systems (NeurIPS)*, 37: 130057–130083.

Douillard, A.; Cord, M.; Ollion, C.; Robert, T.; and Valle, E. 2020. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part XX 16*, 86–102. Springer.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. *arXiv preprint arXiv:2208.01618*.

Gao, S.; Liu, X.; Zeng, B.; Xu, S.; Li, Y.; Luo, X.; Liu, J.; Zhen, X.; and Zhang, B. 2023. Implicit diffusion models for continuous super-resolution. In *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 10021–10030.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Gretton, A.; Borgwardt, K. M.; Rasch, M. J.; Schölkopf, B.; and Smola, A. 2012. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1): 723–773.

Gu, Y.; Wang, X.; Wu, J. Z.; Shi, Y.; Chen, Y.; Fan, Z.; Xiao, W.; Zhao, R.; Chang, S.; Wu, W.; et al. 2023. Mix-of-Show: Decentralized Low-Rank Adaptation for Multi-Concept Customization of Diffusion Models. In *NeurIPS*.

Heng, A.; and Soh, H. 2023. Selective amnesia: A continual learning approach to forgetting in deep generative models. *Advances in Neural Information Processing Systems (NeurIPS)*, 36: 17170–17194.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. LoRA: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Kawar, B.; Zada, S.; Lang, O.; Tov, O.; Chang, H.; Dekel, T.; Mosseri, I.; and Irani, M. 2023. Imagic: Text-based real image editing with diffusion models. In *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 6007–6017.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13): 3521–3526.

Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; and Zhu, J.-Y. 2023. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1931–1941.

Li, Z.; and Hoiem, D. 2017. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12): 2935–2947.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 11461–11471.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *Proc. International Conference on Learning Representations*.

McCloskey, M.; and Cohen, N. J. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, 109–165. Elsevier.

Pham, Q.; Liu, C.; and Hoi, S. 2021. Dualnet: Continual learning, fast and slow. *Advances in Neural Information Processing Systems*, 34: 16131–16144.

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

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv preprint arXiv:2204.06125*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022a. High-resolution image synthesis with latent diffusion models. In *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022b. High-resolution image synthesis with latent diffusion models. In *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2022. DreamBooth: Fine Tuning Text-to-image Diffusion Models for Subject-Driven Generation. *arXiv preprint arxiv:2208.12242*.

Shan, S.; Cryan, J.; Wenger, E.; Zheng, H.; Hanocka, R.; and Zhao, B. Y. 2023. Glaze: Protecting artists from style mimicry by text-to-image models. In *Proc. USENIX Security Symposium*.

Shan, S.; Ding, W.; Passananti, J.; Wu, S.; Zheng, H.; and Zhao, B. Y. 2024. Nightshade: Prompt-specific poisoning attacks on text-to-image generative models. In *2024 IEEE Symposium on Security and Privacy (SP)*, 807–825. IEEE.

Shi, J.; Xiong, W.; Lin, Z.; and Jung, H. J. 2024. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8543–8552.

Smith, J.; Hsu, Y.-C.; Balloch, J.; Shen, Y.; Jin, H.; and Kira, Z. 2021. Always be dreaming: A new approach for data-free class-incremental learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9374–9384.

Smith, J. S.; Hsu, Y.-C.; Kira, Z.; Shen, Y.; and Jin, H. 2024. Continual diffusion with stamina: Stack-and-mask incremental adapters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1744–1754.

Smith, J. S.; Hsu, Y.-C.; Zhang, L.; Hua, T.; Kira, Z.; Shen, Y.; and Jin, H. 2023a. Continual diffusion: Continual customization of text-to-image diffusion with c-lora. *arXiv preprint arXiv:2304.06027*.

Smith, J. S.; Karlinsky, L.; Gutta, V.; Cascante-Bonilla, P.; Kim, D.; Arbelle, A.; Panda, R.; Feris, R.; and Kira, Z. 2023b. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11909–11919.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proc. International Conference on Machine Learning*, 2256–2265.

Van de Ven, G. M.; Siegelmann, H. T.; and Tolias, A. S. 2020. Brain-inspired replay for continual learning with artificial neural networks. *Nature communications*, 11(1): 4069.

Wang, Y.; Huang, Z.; and Hong, X. 2022. S-prompts learning with pre-trained transformers: An occam's razor for domain incremental learning. *Advances in Neural Information Processing Systems*, 35: 5682–5695.

Wang, Z.; Zhang, Z.; Lee, C.-Y.; Zhang, H.; Sun, R.; Ren, X.; Su, G.; Perot, V.; Dy, J.; and Pfister, T. 2022. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 139–149.

Zhai, M.; Chen, L.; Tung, F.; He, J.; Nawhal, M.; and Mori, G. 2019. Lifelong gan: Continual learning for conditional image generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2759–2768.

Zhang, C.; Zhang, C.; Zhang, M.; and Kweon, I. S. 2023. Text-to-image diffusion model in generative ai: A survey. *arXiv preprint arXiv:2303.07909*.

Zhang, L.; and Agrawala, M. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. *arXiv preprint arXiv:2302.05543*.