Disentangling Subject-Irrelevant Elements in Personalized Text-to-Image Diffusion via Filtered Self-distillation

Seunghwan Choi Jooyeol Yun Jeonghoon Park Jaegul Choo
Korea Advanced Institute of Science and Technology (KAIST)
Daejeon, South Korea
{shadow2496, blizzard072, jeonghoon_park, jchoo}@kaist.ac.kr

Abstract

Recent research has unveiled the development of customizing large-scale text-to-image models. These models bind a unique subject desired by a user to a specific token, using the token to generate the subject in various contexts. However, models from previous studies also bind elements unrelated to the subject's identity, such as common backgrounds or poses in the reference images. This often leads to conflicts between the token and the context of text prompts during inference, causing the model to fail to generate both the subject and the prompted context. In this work, we approach this issue from a data scarcity perspective and propose to augment the number of reference images through a novel self-distillation framework. Our framework selects high-quality samples from images generated by a teacher model and uses them in student training. Our framework can be applied to any models that suffer from the conflicts, and we demonstrate that our framework most effectively resolves the issue through comprehensive evaluations.

1. Introduction

Large-scale text-to-image diffusion models [4, 16, 20, 21, 23] excel at generating general subjects (*e.g.*, a dog) given text prompts, but struggle to generate unique subjects (*e.g.*, a user's own dog) that have not been seen during training. To address this limitation, the task of enabling the generation of unique subjects by customizing text-to-image models [10, 11, 13, 14, 22, 28], often referred to as "personalization," has emerged. Specifically, given reference images containing a unique subject and text prompts, the task aims to generate an identical subject in various contexts (*e.g.*, backgrounds, outfits, or poses). To achieve this goal, it is crucial to capture all the visual attributes that define the subject, such as shape, color, and material, while also accurately depicting the context in the prompts.

The task of personalization is generally solved by binding the visual attributes of the subject to a specific token (*e.g.*, [V]). This is done by optimizing the text-to-image models so that the token reconstructs the reference images. Once bound, it is possible to generate the subject by inserting the token into prompts (*e.g.*, "a **[V]** dog swimming").

Unfortunately, since we are given a small (4–6) number of reference images in most cases, existing methods struggle to distinguish between the subject and the elements frequently appearing in the images, such as the background or

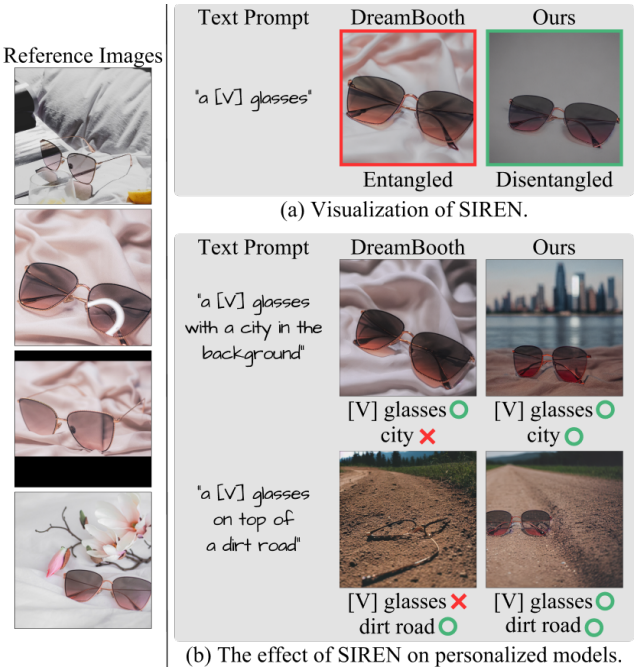

Figure 1. Generation results of DreamBooth [22] and our method for multiple text prompts. (a) The entanglement of subject-irrelevant elements can be observed by examining the generation results for the basic text prompt. DreamBooth depicts the entangled blanket in the image, whereas our method generates only the glasses. (b) While DreamBooth outputs omit the context in the text prompts or the subject, our method generates both without issues.

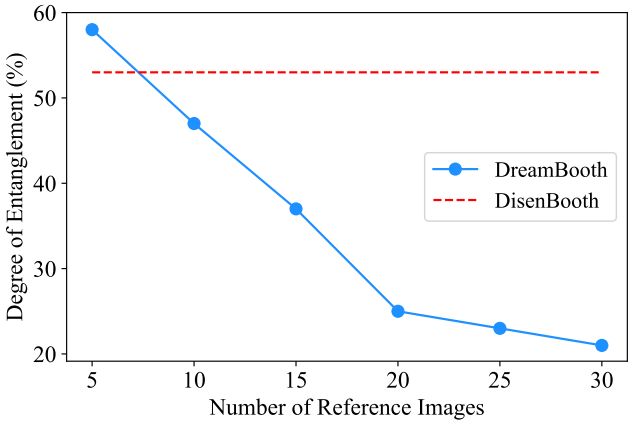

Figure 2. Degrees of SIREN based on the number of reference images. DisenBooth [7] is trained on five images. Details on the images and the degree computation are described in supplement.

the subject's pose. Despite these elements not being part of the subject's identity, they are often mistaken for the subject information and get entangled in the token of the subject. We refer to the phenomenon as Subject-IRrelevant Element eNtanglement (SIREN), which can be observed as in Fig. 1a. When SIREN occurs during personalization, the entangled elements frequently conflict with the context in text prompts during inference. For example, if the background of the beach is entangled in the subject token, it conflicts with the text "in the jungle." This makes it difficult for the model to generate both the subject token and the context in the prompts, resulting in images where only one or the other is depicted. Examples of such images are shown in the DreamBooth [22] results in Fig. 1b. As such, SIREN significantly increases the percentage of the incorrectly generated images, which is detrimental to the personalized models.

Although most personalized models suffer from the entanglement, only a few studies [7, 12] shed light on this matter. They modify the training loss [7] or text prompt [12] to separate subject-irrelevant elements from the subject's identity. However, none of these studies attempt to address the fundamental cause of SIREN: the lack of data. Here, we investigate the effect of the amount of data on SIREN in Fig. 2. We first train the models from the above-mentioned studies [7, 12] on five reference images and draw the lowest degree of SIREN as the red dotted line (*i.e.*, DisenBooth [7]). Then, we train DreamBooth on different numbers of reference images and draw the entanglement degrees as blue lines. Fig. 2 illustrates that adding only five reference images (10 in total) on DreamBooth lowers the entanglement degree more than applying other methods. It also shows that the degree of SIREN constantly decreases as more reference images with different backgrounds or subject poses are added. While we normally only have access to a small number of reference images, this means that if we find a method to obtain images of the subject similar to the reference images, we can most effectively prevent SIREN from occurring. We discover that such images are obtainable from an already SIREN-affected personalized model.

In this work, we tackle the data scarcity issue in the personalization task by proposing a novel self-distillation framework. Our framework is based on the observation that even SIREN-affected models occasionally generate images that accurately depict both the subject and the context in the text prompt, which we refer to as "gold samples." If these samples are used in training alongside reference images, they can show the subject in more diverse environments while conveying consistent visual attributes of the subject. This enables the model to distinguish between the subject and subject-irrelevant elements, thus preventing SIREN.

To obtain the gold samples, we first generate images covering a sufficient variety of backgrounds, subject poses, etc., using a SIREN-affected personalized model (*i.e.*, a teacher model). We then carefully select the samples from these images. Since manually selecting the gold samples is time-consuming, we design and apply an automated filtering mechanism that identifies the samples in a few seconds. The filtered samples are later used in conjunction with reference images to train another model (*i.e.*, a student model).

We demonstrate the superiority of our framework in preventing SIREN through qualitative and quantitative evaluations. Our framework ensures that the model generates both input subjects and prompted contexts without conflicts and achieves the lowest degree of entanglement. Also, we show that our framework is applicable to any personalized model. Since our framework is not dependent on a model's architecture or training objective, we can apply it to any personalized model and improve its subject fidelity and prompt relevance. Lastly, we show that our framework can be applied repeatedly by using the student as a new teacher.

2. Related Work

2.1. Diffusion Models for Text-to-Image Generation

Diffusion models (DMs) [9, 16, 20, 21, 23, 25] are latent variable models that generate samples from a learned data distribution by iteratively denoising Gaussian noise. The training of DMs can be simplified to predicting the denoised image $\mathbf{x}_0$ from a noisy image $\mathbf{x}_t$ at a timestep $t = 1, ..., T$. In a setting that takes text prompts as input, the objective function of a diffusion model $\mathbf{x}_\theta$ is written as

$$\mathbb{E}_{\mathbf{x}_0, y, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), t} ||\mathbf{x}_\theta(\mathbf{x}_t, y, t) - \mathbf{x}_0||_2^2, \quad (1)$$

where y is a text prompt and ϵ is a noise added to $\mathbf{x}_0$ to create $\mathbf{x}_t$. After the training, the model prediction $\mathbf{x}_\theta(\mathbf{x}_t, y, t)$ allows generating a slightly denoised image $\hat{\mathbf{x}}_{t-1}$ from an image $\mathbf{x}_t$ through a sampling step. $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is iteratively passed through the model forward and the sampling step to produce a new denoised image $\hat{\mathbf{x}}_0$ that reflects y.

2.2. Personalization of Text-to-Image Models

Personalization of text-to-image models [3, 10, 11, 13, 14, 22, 28] aims to implant a unique subject into the output domain of a pre-trained text-to-image model, using a few reference images containing the subject. The pioneering approach, Textual Inversion [10], proposes to optimize a token embedding to best reconstruct the reference images. However, a more popular strategy, which is first seen in DreamBooth [22], fine-tunes the entire model to make a fixed subject token reconstruct the images. This strategy, though effective, demands significant computational resources, and various studies have focused on parameter-efficient techniques [11, 13] or encoder-based approaches [14, 28] to reduce the training cost. Nonetheless, existing methods still train the models by reconstructing the small number of reference images that are often limited in diversity. This results in the personalized models binding subject-irrelevant elements to the subject token, leading to outputs that omit the context in inference prompts or the subject itself.

It is until recently that a few studies [7, 12] have attempted to address the Subject-IRrelevant Element eNtanglement (SIREN) by introducing a modified training loss or text prompt. However, as shown in Fig. 2, the fundamental cause of SIREN is the unavailability of diverse training images, and collecting additional data can easily outperform the state-of-the-art approaches. Here, we propose to obtain such data through a novel self-distillation framework and address the core data scarcity issue in personalization.

2.3. Improving Models via Self-distillation

Self-distillation [2, 29, 32, 33] is training a student model using the predictions of an initially trained teacher model with the same architecture. This technique has demonstrated its effectiveness in image understanding tasks such as image classification, especially when the teacher model provides pseudo-labels for unlabeled image samples [29] or assists in mitigating noisy labels within the dataset [8].

In this paper, we extend the concept of self-distillation to generative models, which can be particularly valuable in personalization scenarios where the training data is limited. Specifically, we employ a personalized teacher model to generate diverse samples, which are then carefully selected and added to the training data of a student model. The entire process can be understood as a selective distillation of knowledge from the teacher to the student model.

3. Proposed Method

Given a few (4–6) reference images of an input subject and text prompts, our goal is to generate images that capture the visual attributes (*e.g.*, shape, color, and material) of the subject while also depicting various contexts (*e.g.*, backgrounds, outfits, or poses) described in the prompts. To achieve this, it is crucial to disentangle any subject-irrelevant elements from the subject's token so that the token does not conflict with the text prompts.

In this section, we introduce a self-distillation framework that most effectively alleviates the Subject-IRrelevant Element eNtanglement (SIREN). Our framework first trains a text-to-image model solely on reference images and obtains a SIREN-affected teacher model. Then, the teacher generates numerous images of the subject with diverse backgrounds, poses, etc., from which we select the gold samples. We refer to this stage as "mining." Lastly, the gold samples are used alongside the reference images to train a student model. The student sees the input subject in a wider variety of environments than the teacher, which prevents subject-irrelevant elements from getting entangled in the subject token. All these stages are automated and do not require any human intervention. Fig. 3 illustrates our framework.

3.1. Teacher Training

The first step in our framework is obtaining a personalized teacher model. Since our framework is independent of training objectives, we can apply any existing personalization methods to train the teacher. Among the existing methods, DreamBooth [22] fine-tunes all the parameters of models and captures the visual attributes of subjects with minimal corruption compared to other methods. Therefore, we mainly follow DreamBooth for the teacher training.

We adopt the denoising loss in Eq. (1) as the main loss, where the reference images are used as $\mathbf{x}_0$. The format of the text prompt y is "a [V] [class noun]," where [V] is a rarely-occurring token linked to the input subject and [class noun] is a coarse class descriptor of the subject. For the model's architecture, we use Stable Diffusion [1], a publicly available text-to-image diffusion model. More details of the teacher training are described in the supplement.

3.2. Mining Gold Samples

We now explain the process of acquiring the gold samples from the personalized teacher model. To ensure that the samples cover a sufficient variety of backgrounds, subject poses, etc., we first construct a set of text prompts describing diverse environments. These prompts are referred to as "mining prompts." Then, we input the prompts into the teacher and generate numerous images of the input subject. Since the teacher is not immune to SIREN, most generated images would be missing the context in the prompts or the subject itself. We filter out these images by devising a filtering mechanism to evaluate the quality of each image in an automated manner and acquire the gold samples we seek.

Constructing Mining Prompts. Mining prompts should describe a wide variety of subject-irrelevant elements that are likely to be entangled in the subject token. These ele-

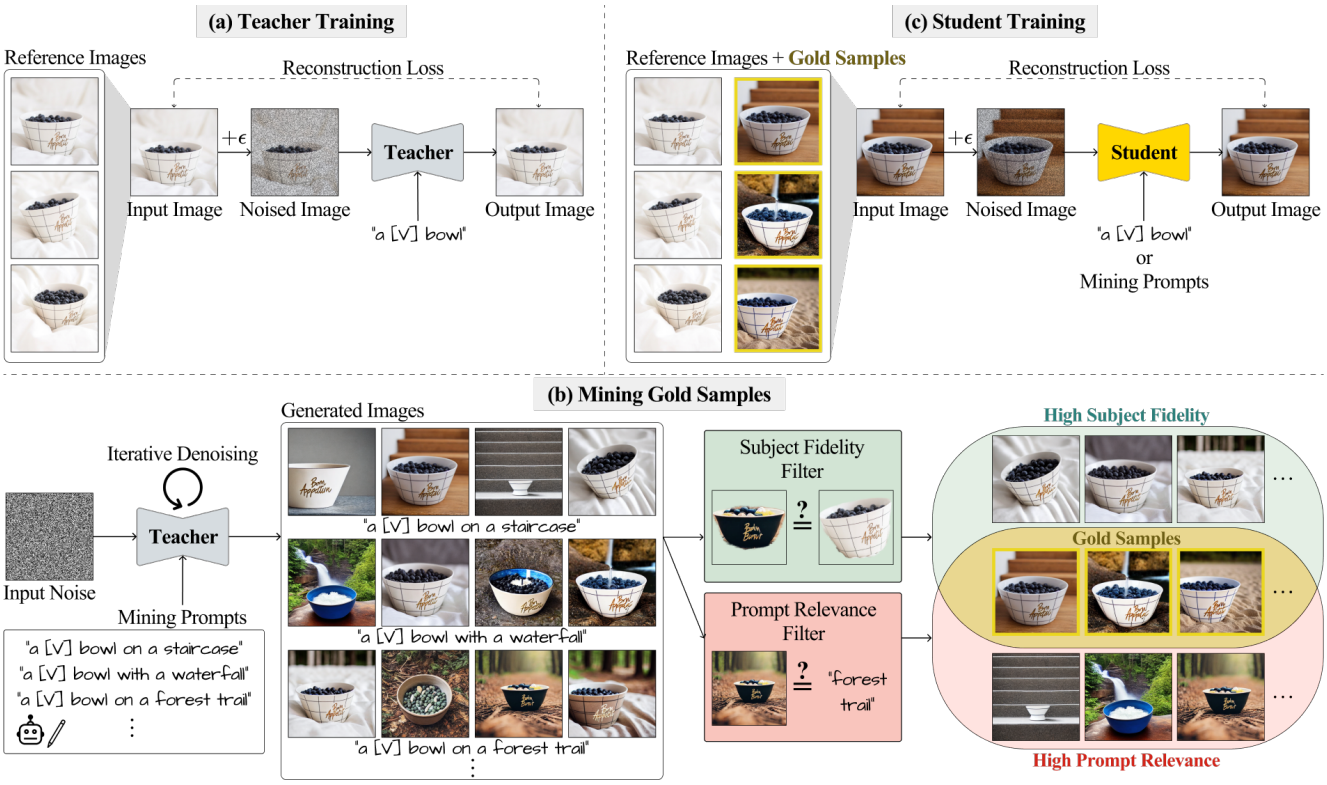

Figure 3. An overview of the framework. (a) We train the teacher model solely on the reference images. (b) The trained teacher generates numerous images from the pre-constructed mining prompts, and we select the gold samples from those images by evaluating the quality of each image in an automated manner. (c) The selected gold samples are used in student training alongside the reference images.

ments include backgrounds, outfits, poses, and more. Constructing such prompts would have traditionally required manual crafting by humans. However, recent advancements in large language models (LLMs) [5, 18, 26, 27] have made this type of job significantly convenient. Thus, we obtain the mining prompts utilizing an LLM, in our case ChatGPT [17]. Specifically, we provide ChatGPT with a few exemplar prompts containing the subject-irrelevant elements and retrieve new sentences in the same format. The retrieved sentences are subsequently templated and used as mining prompts for all input subjects. A comprehensive list of the mining prompts is reported in the supplementary material.

Evaluation of Generated Images. There are two major evaluation criteria for assessing the quality of generated images in the personalization task: subject fidelity and prompt relevance. Subject fidelity refers to the degree of resemblance between a generated subject and the input subject. It is measured by the average pairwise cosine similarity between the embeddings of the generated image and the reference images. There are CLIP-I metric that uses CLIP [19] embeddings, and DINO metric that utilizes ViT-S/16 DINO [6] embeddings. Prompt relevance gauges the degree to which the prompted context is depicted in the generated image. This is measured by the CLIP-T metric, which calculates the average cosine similarity between the CLIP embeddings of the generated image and the prompt.

Here, we aim to assess the quality of the generated images using improved metrics. One of these metrics is CLIP-MI. It is a variation of CILP-I that extracts the foreground corresponding to the subject area from the images before computing the cosine similarity. This has the advantage of excluding the background, one of the major contexts, from affecting the similarity. For prompt relevance, most vision-language models [15, 24, 31], including CLIP, have the drawback of lacking compositional understanding. To address this, Yuksekgonul *et al*. [30] proposes NegCLIP, a model that improves the compositional understanding of CLIP. We use NegCLIP to compare the generated image and the prompt, namely NegCLIP-T. A detailed comparison of CLIP-MI, NegCLIP-T, and the existing evaluation metrics is available in the supplementary material.

Selecting Gold Samples. Utilizing the newly proposed metrics, we select gold samples from the generated images. The idea is simple yet effective. Based on each metric, the images can be divided into two groups: those whose scores belong to the top k% of all the generated images and those that do not. Therefore, we obtain a group of images with high subject fidelity based on CLIP-MI and one with high

prompt relevance based on NegCLIP-T. The intersection of the two groups consists of the generated images with both high subject fidelity and prompt relevance, corresponding to the gold samples we seek. The high quality of the samples allows them to be regarded as pseudo-reference images and thus are able to be used in student training.

3.3. Student Training

Now that we have the gold samples, we can train the student model. The model has the same architecture as the teacher, and the training procedure is identical. One difference is the number of training images, as the model uses not only the reference images but also the gold samples. This exposes the model to a wider variety of environments than the teacher, preventing subject-irrelevant elements from being entangled in the token of the input subject. Additionally, for each gold sample, we use the mining prompt used in its generation as the training prompt. Using text prompts that explicitly mention subject-irrelevant elements has the effect of further separating the elements from the subject token.

4. Experiments

Dataset. We train the teacher and student models on the DreamBooth dataset [22]. The dataset comprises 21 objects and 9 live subjects/pets with 4–6 reference images per subject. It also includes 25 evaluation prompts covering recontextualization, accessorization, and property modification.

Evaluation Metrics. We evaluate our method with the newly proposed CLIP-MI and NegCLIP-T [30] metrics. CLIP-MI measures subject fidelity, while NegCLIP-T measures prompt relevance. Also, we can quantify the degree of SIREN by the proportion of images that had conflicts during generation. The conflict-occurred images are characterized by excessively lacking representations of the subject or the context in evaluation prompts. We classify an image with a maximum CLIP-MI or NegCLIP-T score below the minimum threshold as a conflict-occurred image. We compute the proportion of such images and call it the "conflict rate."

4.1. Mining Results

The mining stage of our framework successfully yields gold samples in an automated manner, and the average number of the yielded samples per subject is 17. Fig. 4 shows snippet of these samples. We can see that the samples depict sufficiently diverse contexts thanks to the mining prompts constructed to contain different backgrounds, poses, and outfits with the assistance of the LLM. Furthermore, the figure demonstrates that our filtering mechanism carefully selects only high-quality images as gold samples. The selected samples capture almost all the visual attributes of the subjects while accurately depicting the context of the mining prompts. We provide full mining results in the supplement.

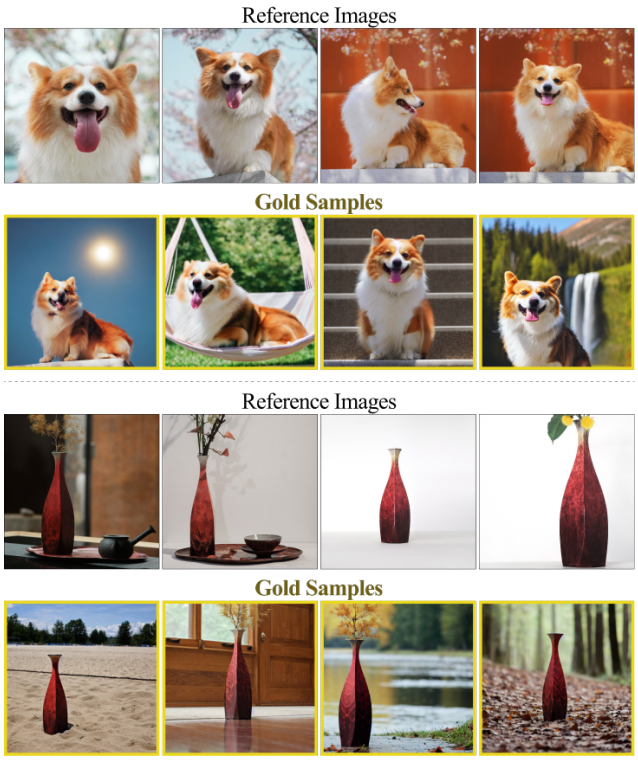

Figure 4. Mining results. The gold samples capture almost all visual attributes of subjects and depict sufficiently diverse contexts.

4.2. Comparisons

We compare our method with DreamBooth [22], Custom Diffusion [13], BLIP-Diffusion [14], DisenBooth [7], and SID [12]. All the methods are reproduced using the official codes or with the implementation details reported in each corresponding paper. Fig. 5 shows qualitative comparison. The figure, in general, demonstrates that while the compared methods struggle to generate the subjects and the prompted contexts simultaneously, our method generates both without issues. Also, if we compare our method with DreamBooth, which can be considered as our teacher model, we can visually see the difference in the degree of SIREN depending on whether the gold samples are used for training or not. Since DreamBooth is trained solely on the reference images, the subject token frequently conflicts with the evaluation prompts, and the model mostly replicates the reference images or generates different subjects from the input. On the other hand, due to the use of gold samples during training, our method generates images with both high subject fidelity and prompt relevance. It is also noteworthy that DisenBooth and SID still suffer from SIREN despite the modified training loss or text prompt they propose.

For quantitative evaluation, we generate 50 images per evaluation prompt, resulting in 1,250 images per subject. We evaluate these images with the evaluation metrics and

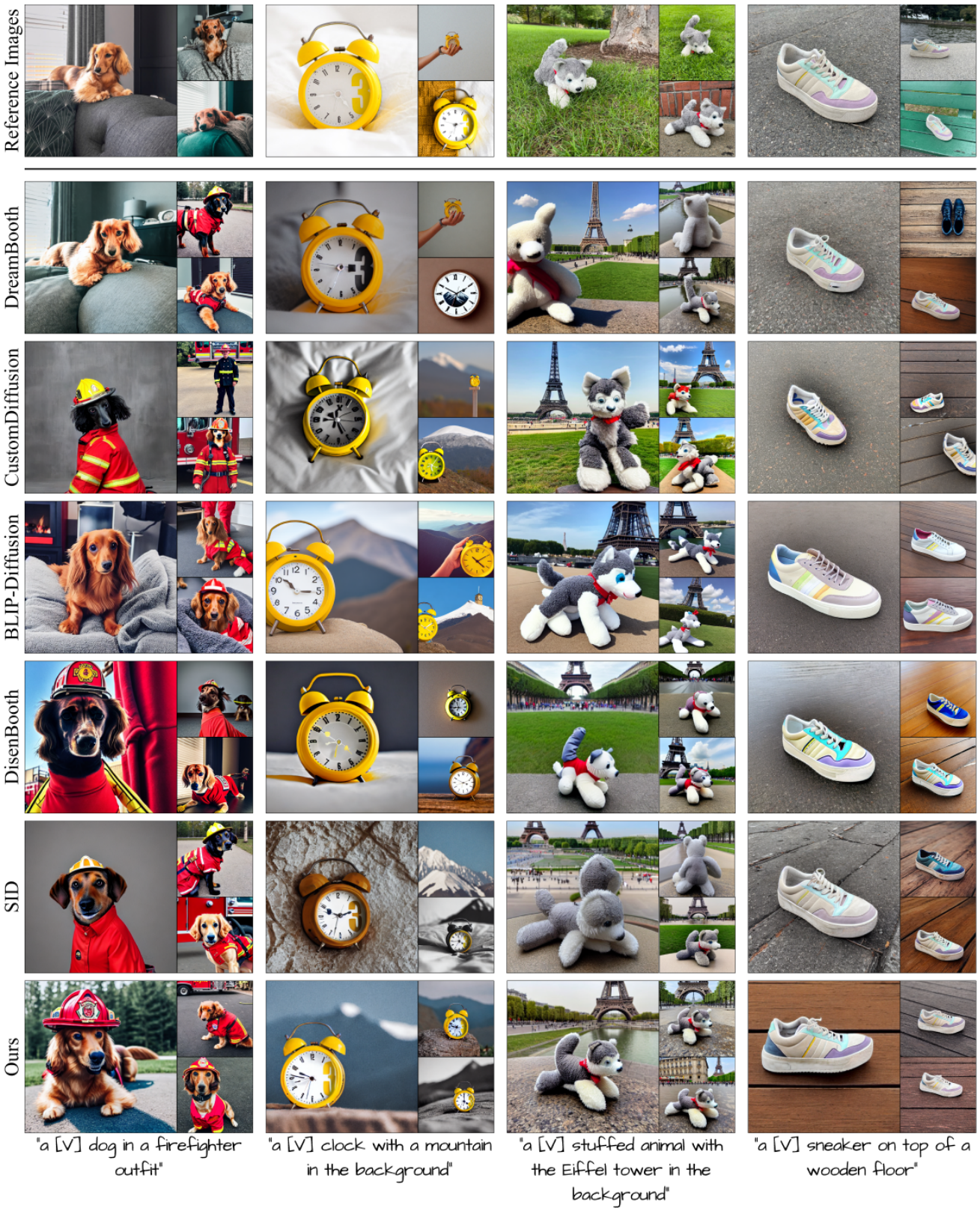

Figure 5. Qualitative comparison with existing personalization methods [7, 12–14, 22]. While the existing methods struggle to generate the subjects and the prompted contexts simultaneously, our method excels in generating both without issues.

Method	CLIP-MI$_\uparrow$	NegCLIP-T$_\uparrow$	Conflict Rate$_\downarrow$
Real Images	0.947	N/A	N/A
DreamBooth [22]	0.851	0.306	24.0%
Custom Diffusion [13]	0.824	0.309	22.4%
BLIP-Diffusion [14]	0.852	0.287	38.6%
DisenBooth [7]	0.834	0.305	23.3%
SID [12]	0.777	**0.314**	24.3%
Ours	**0.872**	0.310	**19.4%**

Table 1. Quantitative comparison with existing methods. Ours records the highest CLIP-MI score and the lowest conflict rate.

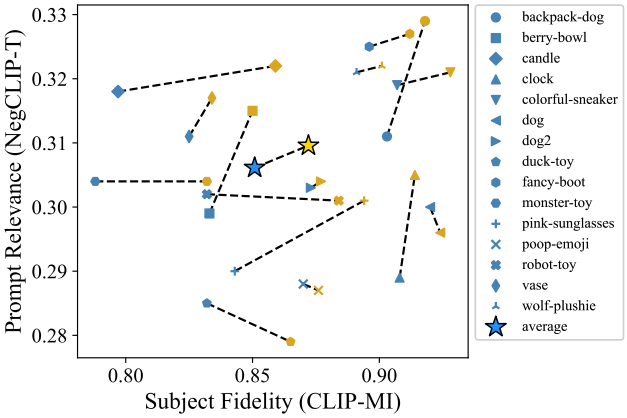

Figure 6. Per-subject scores of DreamBooth (blue) and our method (yellow). Our proposed self-distillation framework significantly improves subject fidelity for each subject. The framework also enhances prompt relevance on average.

report the scores in Table 1. The table demonstrates that our method achieves the highest subject fidelity and comparable prompt relevance. While SID scores highest on NegCLIP-T, the method scores lowest on CLIP-MI due to deficient learning of subject information. Additionally, our method records the lowest conflict rate, which aligns with the qualitative comparison. In Fig. 6, we illustrate the subject-wise scores of the teacher model, DreamBooth, and our method. The figure shows that the proposed self-distillation framework significantly improves subject fidelity for each subject and enhances prompt relevance on average.

4.3. Ablation Studies

We also conduct ablation studies for two setups. One setup uses "a [V] [class noun]" as the training prompt for the gold samples instead of the detailed mining prompts. While this setup still achieves a lower conflict rate than DreamBooth, as shown in Table 2, Fig. 7 shows that the model often generates inaccurate contexts or omits several subject details. This demonstrates that using detailed prompts that explicitly mention subject-irrelevant elements helps further separate the elements from the subject token.

Ablation Setup	CLIP-MI$_\uparrow$	NegCLIP-T$_\uparrow$	Conflict Rate$_\downarrow$
DreamBooth [22]	0.851	0.306	24.0%
Ours	**0.872**	**0.310**	**19.4%**
Ours w/o detailed prompts	0.852	0.308	22.2%
DB w/ random backgrounds	0.843	0.308	22.8%

Table 2. Quantitative comparison with ablation setups. All setups outperform DreamBooth, but our full method is the best.

Figure 7. Qualitative comparison with ablation setups. Images are generated with the following prompts from left to right: "*a [V] toy in the snow*," "*a [V] toy on top of pink fabric*," "*a [V] toy on the beach*," and "*a [V] toy on top of green grass with sunflowers around it*." Our full method best captures the subject details and the prompted contexts, naturally placing the subject within images.

Another setup involves replacing the backgrounds of the reference images with other random backgrounds during the training of the teacher model. This can be viewed as training the student model on reference images with background augmentation instead of the gold samples. One might think that these augmented images would serve as pseudo-reference images, but Table 2 shows that these images are not as effective as the gold samples in alleviating SIREN. Moreover, the augmented images induce a different form of entanglement. Since the subjects are artificially composited with the backgrounds in these images, the model struggles to naturally place the subjects within images. Refer to Fig. 7 for the generation results.

Method	CLIP-MI$_\uparrow$	NegCLIP-T$_\uparrow$	Conflict Rate$_\downarrow$
DisenBooth	0.834	0.305	23.3%
DisenBooth + Ours	**0.836**	**0.312**	**17.8%**

Table 3. Self-distillation on DisenBooth. Our method reduces the conflict rate, improving subject fidelity and prompt relevance.

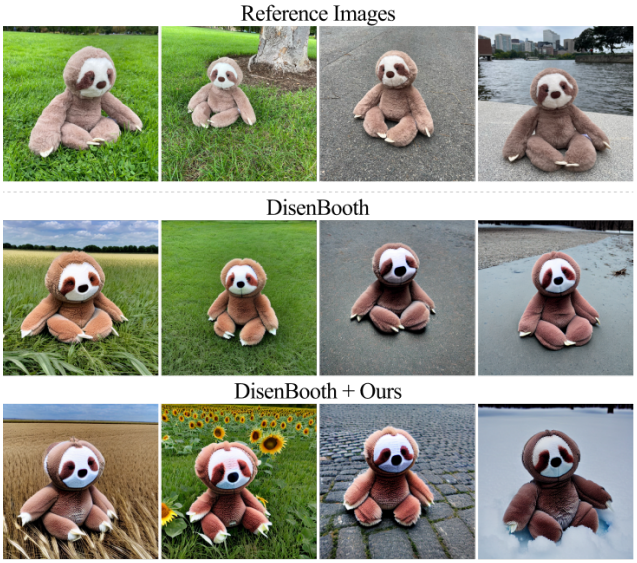

Figure 8. Self-distillation on DisenBooth. Images are generated with the following prompts from left to right: "*a [V] stuffed animal with a wheat field in the background*," "*a [V] stuffed animal on top of green grass with sunflowers around it*," "*a [V] stuffed animal on a cobblestone street*," and "*a [V] stuffed animal in the snow*."

4.4. Additional Experiments

Self-distillation on DisenBooth. Our method is applicable to any personalization model. To demonstrate this, we report the results of applying our method to DisenBooth in Table 3 and Fig. 8. Similar to when DreamBooth is used as the teacher model, our method significantly reduces the conflict rate, leading to improvements in subject fidelity and prompt relevance.

Repeating Self-distillation. Our framework allows using the student model as a new teacher. To demonstrate this, we report the results for a subject with a high collision rate, pink-sunglasses, in Table 4. The results highlight the unique strength of our approach, as the performance progressively improves as the iterations increase.

5. Limitations

While the filtering mechanism of our self-distillation framework selects only high-quality images as the gold samples, there is still a possibility that generated images with distorted subjects are chosen. To prevent this at the expense of the yield of the gold samples, we can increase the threshold of the subject fidelity filter. Additionally, the generation capability of our student model for complex prompts, such as compositional prompts, is limited to that of Stable Diffusion. Fortunately, our framework is not dependent on Stable Diffusion, but to improve such capability, changing the architecture is required nonetheless.

Metric	DreamBooth	Ours	Ours×2	Ours×3
CLIP-MI$_\uparrow$	0.843	0.894	0.899	**0.906**
NegCLIP-T$_\uparrow$	0.290	0.301	0.304	**0.306**
Conflict Rate$_\downarrow$	38.5%	23.4%	20.3%	**15.8%**

Table 4. Repeated self-distillation results on the pick-sunglasses subject. The performance improves as the iteration increases.

6. Conclusion

In this paper, we introduce a method designed to mitigate SIREN, a common issue observed in most personalized text-to-image models. We demonstrate that SIREN significantly harms the user experience, often resulting in samples that omit the context in inference prompts or the subject. To tackle this challenge, we propose a self-distillation framework where a trained teacher model provides the student model with diverse samples encompassing subjects in various environments.

Based on our analysis that even a SIREN-affected personalized model occasionally generates high-quality samples, we implement a filtering process to select these gold samples from a pool of images generated by a teacher model. By training the student model with the gold samples and the original reference images, our approach effectively mitigates the SIREN, enabling the model to accurately depict subjects in various contexts. Our approach establishes a virtuous cycle for improving the performance of personalized models and is easily extendable to various personalization approaches.

Acknowledgements

We thank Yunjey Choi, Junho Kim, and Hyunsu Kim for their valuable input that helped advance this work. This work was partly done by Seunghwan Choi during the NAVER AI Lab internship. This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (No.RS-2019-II190075 Artificial Intelligence Graduate School Program(KAIST)) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2022R1A2B5B02001913).

References

[1] Stability AI. Stable diffusion. https://huggingface.co/runwayml/stable-diffusion-v1-5, 2022. [Online; accessed 3-November-2023]. 3

[2] Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. In *Proc. the International Conference on Learning Representations (ICLR)*, 2022. 3

[3] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. *arXiv preprint arXiv:2305.16311*, 2023. 3

[4] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. https://cdn.openai.com/papers/dall-e-3.pdf, 2023. [Online; accessed 4-December-2024]. 1

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Proc. the Advances in Neural Information Processing Systems (NeurIPS)*, 33:1877–1901, 2020. 4

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proc. of the IEEE international conference on computer vision (ICCV)*, pages 9650–9660, 2021. 4

[7] Hong Chen, Yipeng Zhang, Simin Wu, Xin Wang, Xuguang Duan, Yuwei Zhou, and Wenwu Zhu. Disenbooth: Identity-preserving disentangled tuning for subject-driven text-to-image generation. In *Proc. the International Conference on Learning Representations (ICLR)*, 2023. 2, 3, 5, 6, 7

[8] Rudrajit Das and Sujay Sanghavi. Understanding self-distillation in the presence of label noise. *Proc. the International Conference on Machine Learning (ICML)*, 2023. 3

[9] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Proc. the Advances in Neural Information Processing Systems (NeurIPS)*, 34:8780–8794, 2021. 2

[10] 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. *arXiv preprint arXiv:2208.01618*, 2022. 1, 3

[11] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. *arXiv preprint arXiv:2303.11305*, 2023. 1, 3

[12] Jimyeong Kim, Jungwon Park, and Wonjong Rhee. Selectively informative description can reduce undesired embedding entanglements in text-to-image personalization. In *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 8312–8322, 2024. 2, 3, 5, 6, 7

[13] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 1931–1941, 2023. 1, 3, 5, 6, 7

[14] Dongxu Li, Junnan Li, and Steven CH Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *arXiv preprint arXiv:2305.14720*, 2023. 1, 3, 5, 6, 7

[15] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proc. the International Conference on Machine Learning (ICML)*, pages 12888–12900, 2022. 4

[16] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 1, 2

[17] OpenAI. Chatgpt. https://chat.openai.com/, 2022. [Online; accessed 4-December-2024]. 4

[18] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Proc. the Advances in Neural Information Processing Systems (NeurIPS)*, 35:27730–27744, 2022. 4

[19] 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 *Proc. the International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021. 4

[20] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1, 2

[21] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 10684–10695, 2022. 1, 2

[22] 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 *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 22500–22510, 2023. 1, 2, 3, 5, 6, 7

[23] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Proc. the Advances in Neural Information Processing Systems (NeurIPS)*, 35:36479–36494, 2022. 1, 2

[24] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proc. of the IEEE conference on com-*

puter vision and pattern recognition (CVPR), pages 15638–15650, 2022. 4

[25] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proc. the International Conference on Machine Learning (ICML)*, pages 2256–2265, 2015. 2

[26] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 4

[27] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*, 2022. 4

[28] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*, 2023. 1, 3

[29] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 10687–10698, 2020. 3

[30] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *Proc. the International Conference on Learning Representations (ICLR)*, 2022. 4, 5

[31] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276*, 2021. 4

[32] Linfeng Zhang, Chenglong Bao, and Kaisheng Ma. Self-distillation: Towards efficient and compact neural networks. *The IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(8):4388–4403, 2021. 3

[33] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proc. of the IEEE international conference on computer vision (ICCV)*, pages 3713–3722, 2019. 3