

# APT: Adaptive Personalized Training for Diffusion Models with Limited Data

JungWoo Chae<sup>1\*</sup> Jiyo Kim<sup>1\*</sup> JaeWoong Choi<sup>1</sup> Kyungul Kim<sup>1</sup> Sangheum Hwang<sup>2†</sup>

<sup>1</sup>LG CNS AI Research <sup>2</sup>Department of Data Science, Seoul National University of Science and Technology

{cjwoolgcn, jiyo.onkim, jaewoong.choi, kyungul.kim }@lgcns.com

shwang@ds.seoultech.ac.kr

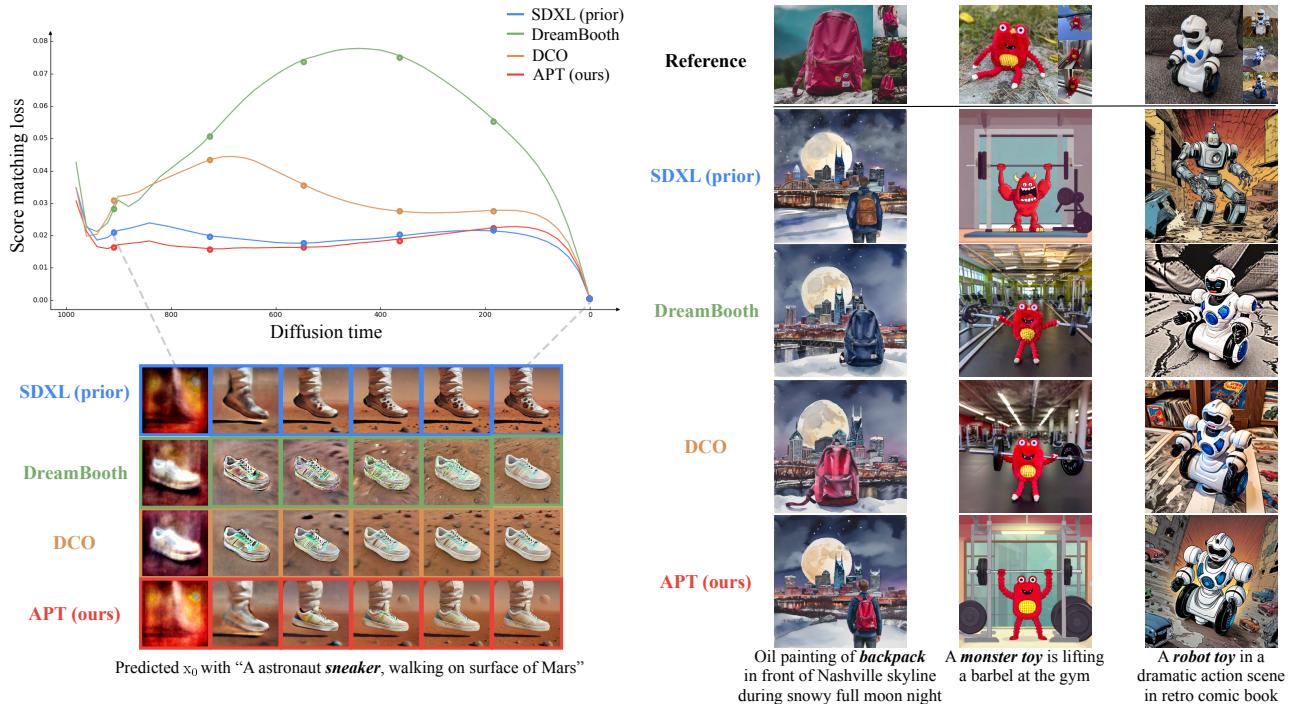


Figure 1. Given a few reference images, APT personalizes diffusion models with less overfitting: (Left) By comparing diffusion trajectories using the score matching loss [6], we observe that our method maintains the original denoising path. The predicted  $x_0$  images from APT closely resemble SDXL (prior) during early steps, preserving the overall layout and scene context. (Right) APT effectively incorporates contextual elements from the prior, such as generating a backpack with a person without explicitly mentioning “person” and preserves stylistic elements like comic book aesthetics. In contrast, other methods either focus excessively on reference images or fail to maintain the prior’s style. This demonstrates that APT successfully maintains the pretrained model’s capabilities for text alignment and stylization.

## Abstract

Personalizing diffusion models using limited data presents significant challenges, including overfitting, loss of prior knowledge, and degradation of text alignment. Overfitting leads to shifts in the noise prediction distribution, disrupting the denoising trajectory and causing the model to lose semantic coherence. In this paper, we propose Adap-

tive Personalized Training (APT), a novel framework that mitigates overfitting by employing adaptive training strategies and regularizing the model’s internal representations during fine-tuning. APT consists of three key components: (1) **Adaptive Training Adjustment**, which introduces an overfitting indicator to detect the degree of overfitting at each time step bin and applies adaptive data augmentation and adaptive loss weighting based on this indicator; (2) **Representation Stabilization**, which regularizes the mean and variance of intermediate feature maps to prevent exces-

\*Equal contribution

†Corresponding author

sive shifts in noise prediction; and (3) **Attention Alignment for Prior Knowledge Preservation**, which aligns the cross-attention maps of the fine-tuned model with those of the pre-trained model to maintain prior knowledge and semantic coherence. Through extensive experiments, we demonstrate that APT effectively mitigates overfitting, preserves prior knowledge, and outperforms existing methods in generating high-quality, diverse images with limited reference data.

## 1. Introduction

The advent of diffusion models has significantly advanced the field of generative modeling, enabling the synthesis of diverse and high-quality images [13, 26, 33]. Personalization techniques, such as DreamBooth [27] and Textual Inversion [8], have further enhanced these models by enabling subject-driven generation tailored to specific user needs. Such advancements have broad applications, from artistic content creation to specialized data augmentation in machine learning tasks [37, 38]. However, personalizing diffusion models using limited data presents significant challenges. One critical issue is overfitting, which causes excessive shifts in the noise prediction distribution, disrupting the denoising trajectory of the pretrained diffusion model [17, 27]. These shifts lead to the loss of prior knowledge, degradation of text alignment, and a reduced ability of the model to generalize to unseen prompts.

Overfitting may cause the model to memorize spatial layouts, resulting in generated images with overly similar compositions, or to over-memorize textures, leading to poor stylization and a lack of diversity in response to different prompts. Moreover, overfitting can lead to the loss of prior knowledge, causing the model to generate images that do not accurately reflect the desired concept or context. For example, when generating an image with “*a photo of a backpack*”, pretrained diffusion models may naturally include a person carrying the backpack, leveraging prior knowledge about common contexts. However, after fine-tuning with limited data, the model may lose this prior knowledge, resulting in images of only the backpack without a person. This loss of prior knowledge is accompanied by changes in the cross-attention maps, which further degrade the quality and coherence of the generated images.

Existing methods [9, 17, 21, 32, 38] have addressed these challenges through various regularization techniques and novel fine-tuning approaches. Techniques that constrain attention using masks [3] often require additional annotations and may not align effectively with the soft attention distributions of the model. Furthermore, prior preservation techniques that incorporate additional data [27] fine-tune the model by combining subject and auxiliary images but often suffer from overfitting. This can disrupt the original denoising trajectory, resulting in overfitting to auxiliary datasets

and reduced generalization in generation quality [17].

In this work, we propose **Adaptive Personalized Training (APT)**, a novel framework that addresses these challenges by mitigating overfitting with adaptive training strategies, regularizing the internal representations of the model during fine-tuning, and preserving prior knowledge. Specifically, our method consists of three key components:

1. **Adaptive Training Adjustment:** We introduce an overfitting indicator to detect the degree of overfitting and apply adaptive data augmentation and loss weighting based on this indicator. This approach addresses the varying influence of diffusion model parameters across different time steps due to the beta scheduling, effectively mitigating overfitting and adjusting the training dynamics.
2. **Representation Stabilization:** We regularize the significant shifts in the noise prediction  $\epsilon$  caused by overfitting by constraining the mean and variance of the intermediate feature maps. This helps preserve the statistical properties of the representations of the pretrained model.
3. **Attention Alignment for Prior Knowledge Preservation:** To maintain the prior knowledge in the text embeddings, we propose regularizing the cross-attention maps. By aligning the attention distributions of the fine-tuned model with those of the pretrained model, we ensure that the model retains semantic coherence.

Our contributions can be summarized as follows:

- We introduce **Adaptive Personalized Training (APT)**, a novel method that addresses overfitting and the loss of prior knowledge in diffusion model personalization with limited data. APT incorporates adaptive training adjustments, representation stabilization, and attention alignment to mitigate overfitting and preserve prior knowledge during fine-tuning.
- Through extensive experiments, we demonstrate that APT outperforms existing methods in preserving the text alignment ability and prior knowledge of pretrained models, while generating high-quality and diverse personalized images.

Our method provides a cohesive solution that addresses both the varying influence of model parameters across time steps and the internal representation shifts that arise during fine-tuning with limited data. By mitigating overfitting and preserving prior knowledge, we enable the model to generalize better to unseen prompts while accurately capturing the desired concepts from the reference data.

## 2. Related Work

**Text-to-Image Personalization** Recent advances in diffusion models have enabled high-quality image synthesis through large-scale datasets and advanced architectures [4, 19, 25, 26, 28, 29, 36], with techniques like classifier-free guidance [7, 12, 31] enhancing text alignment.

Personalization of text-to-image models adapts pre-trained models to represent new concepts based on user-provided images. Key methods include DreamBooth [27], which fine-tunes the entire model for high fidelity, and Textual Inversion [8], which optimizes textual embeddings without altering model weights for efficiency. Parameter-efficient fine-tuning methods, such as LoRA [14], Custom Diffusion [16], and Svdiff [9], update only a small subset of parameters to reduce resource demands while maintaining quality. Recent advances like P+ [34] and NeTI [2] expand textual conditioning spaces, enabling greater control and expressiveness without full model fine-tuning, achieving faster convergence and improved editability.

**Regularization in T2I Personalization** Maintaining the prior knowledge of pretrained models during personalization is essential to prevent concept drift. Techniques like the prior preservation loss in DreamBooth [27] limit deviations from the original distribution but struggle with limited data, leading to inconsistencies and undesirable shifts [17]. Recent methods like DCO [17] address this by directly regularizing the KL divergence, while Attention Regularization [21] improves identity preservation through refined cross-attention maps. However, these methods primarily target specific components, such as cross-attention, and fail to fully preserve the pretrained model’s diffusion trajectories, affecting text alignment and diversity.

To overcome these limitations, we propose a method that regularizes not only cross-attention but also self-attention, as well as intermediate representations such as U-Net outputs. By aligning these components along the diffusion process, our approach preserves the pretrained model’s original capabilities while enabling accurate personalization.

**Adaptive Data Augmentation in Generative Models** Overfitting in generative models trained on limited data is a critical challenge. StyleGAN ADA [15] addresses this in GANs by applying augmentations adaptively based on the degree of overfitting, stabilizing training without modifying loss functions or network architectures. Improved Consistency Regularization [39] similarly enhances GANs by enforcing consistency on the discriminator.

While these methods target GANs, adaptive augmentation to mitigate overfitting is also relevant for diffusion models. In our work, we introduce an adaptive augmentation strategy based on the proposed overfitting indicator, dynamically adjusting augmentation strength to prevent overfitting in personalization.

### 3. Method

Personalizing diffusion models with limited reference data introduces significant challenges, such as overfitting, loss of prior knowledge, and degradation of text alignment

[8, 17, 27]. To address these issues, we propose **Adaptive Personalized Training (APT)**, a method focused on mitigating overfitting through adaptive training strategies (Section 3.1), stabilizing the model’s internal representations during fine-tuning (Section 3.2), and preserving prior knowledge (Section 3.3). An overview of our method is illustrated in Figure 2.

#### 3.1. Adaptive Training Adjustment

Fine-tuning diffusion models on limited data can lead to overfitting, where the model excessively memorizes the training data. Due to the beta scheduling in diffusion models, the loss magnitude varies greatly across different time steps, affecting model updates differently at each step. This overfitting causes significant shifts in noise prediction  $\epsilon$ , disrupting the denoising trajectory of the pretrained diffusion model, as observed in Figure 1. This disruption results in the degradation of text alignment and loss of prior knowledge. Therefore, it is necessary to detect and mitigate overfitting by introducing an overfitting indicator and applying adaptive strategies based on it. By adjusting the training dynamics adaptively, we aim to mitigate overfitting and maintain the integrity of the denoising trajectory.

**Adaptive Overfitting Indicator** We introduce an adaptive overfitting indicator  $\gamma_t$  to quantify the degree of overfitting during fine-tuning:

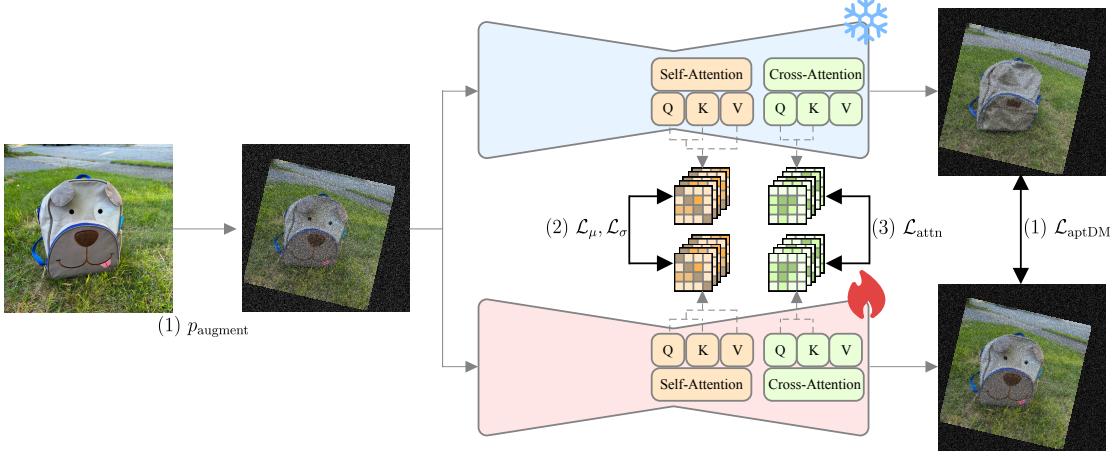
$$\gamma_t = 1 - e^{-T(\text{EMA}_t[\mathcal{L}_{\text{DM}}^\phi] - \text{EMA}_t[\mathcal{L}_{\text{DM}}^\theta])}, \quad (1)$$

where  $T$  is the total number of denoising steps,  $\mathcal{L}_{\text{DM}}^\phi$  and  $\mathcal{L}_{\text{DM}}^\theta$  are the denoising losses of the pretrained model  $\phi$  and the fine-tuned model  $\theta$ , respectively. The  $\text{EMA}_t$  denotes the exponential moving average computed at the specific time step bin  $t$  to reduce fluctuations due to noise and data variance. This formulation ensures that  $\gamma_t = 0$  when there is no overfitting and  $\gamma_t \rightarrow 1$  when overfitting is maximal. In practice, we divide the total diffusion steps into  $B$  bins (e.g.,  $B = 10$  bins of 100 steps each for a total of 1000 steps). The overfitting indicator  $\gamma_t$  is computed separately for each bin  $t$ , capturing the degree of overfitting at different noise levels.

**Adaptive Data Augmentation** We use  $\gamma_t$  as the data augmentation probability, clamping it within a predefined range:

$$p_{\text{augment}} = \text{clamp}(\gamma_t, 0, p_{\max}), \quad (2)$$

where  $p_{\max}$  is the maximum augmentation probability. As shown in Figure 1, the personalized model  $\theta$  tends to memorize spatial configurations from early denoising steps, leading to positional overfitting. To disrupt this memorization, we apply affine transformations as data augmentation. By adjusting the probability of applying data augmentation based on  $\gamma_t$ , we aim to mitigate spatial overfitting.



**Figure 2. Overview of Our Proposed Method (APT).** Illustration of the three key components: (1) Adaptive Training Adjustment with adaptive data augmentation ( $p_{\text{augment}}$ ) and loss weighting ( $\mathcal{L}_{\text{aptdm}}$ ) to mitigate overfitting; (2) Representation Stabilization through regularizing intermediate feature maps to stabilize the noise trajectory ( $\mathcal{L}_\mu, \mathcal{L}_\sigma$ ); (3) Attention Alignment to preserve prior knowledge by regularizing the cross-attention maps ( $\mathcal{L}_{\text{attn}}$ ).

**Adaptive Loss Weighting** In addition to adaptive data augmentation, we adjust the loss weighting adaptively according to the degree of overfitting. We design a weighting scheme that scales the loss for each time step bin based on the degree of overfitting<sup>1</sup>:

$$\mathcal{L}_{\text{aptdm}} = (1 - \gamma_t) \mathcal{L}_{\text{DM}}, \quad (3)$$

where  $\gamma_t$  is the overfitting indicator for time step bin  $t$ , and  $\mathcal{L}_{\text{DM}}$  is the denoising loss. By scaling the loss with  $(1 - \gamma_t)$ , we reduce its impact for time steps where overfitting is detected, effectively rebalancing the training dynamics and mitigating overfitting.

### 3.2. Representation Stabilization

To prevent the denoising trajectory of the fine-tuned model from deviating excessively from the original (i.e., the pre-trained model’s trajectory), it is necessary to regularize these shifts by stabilizing the intermediate representations.

We apply regularization to the mean and variance of the intermediate feature maps of the model to preserve the statistical properties of the representations of the pre-trained model. Let  $\mathbf{h}_\theta^{(l)}$  and  $\mathbf{h}_\phi^{(l)}$  denote the activations at layer  $l$  for the fine-tuned model  $\theta$  and the pre-trained model  $\phi$ , respectively. We define the representation regularization losses as:

$$\mathcal{L}_\mu = \sum_l^{\text{layers}} \left| \mu \left( \mathbf{h}_\theta^{(l)}(x_t; c^*, t) \right) - \mu \left( \mathbf{h}_\phi^{(l)}(x_t; c, t) \right) \right|_2^2, \quad (4)$$

$$\mathcal{L}_\sigma = \sum_l^{\text{layers}} \left| \sigma \left( \mathbf{h}_\theta^{(l)}(x_t; c^*, t) \right) - \sigma \left( \mathbf{h}_\phi^{(l)}(x_t; c, t) \right) \right|_2^2, \quad (5)$$

<sup>1</sup>The motivation for adaptive loss weighting is described in Supplementary Material B.3.

where  $c^*$  is the conditioning information including the identifier (e.g., “V\*”) while  $c$  is the conditioning information with the class token (e.g., “dog”).  $\mu(\cdot)$  and  $\sigma(\cdot)$  compute the mean and standard deviation of activations, respectively. By regularizing these statistics, we limit excessive shifts in the distribution of the intermediate representations, preserving prior knowledge and improving text alignment.

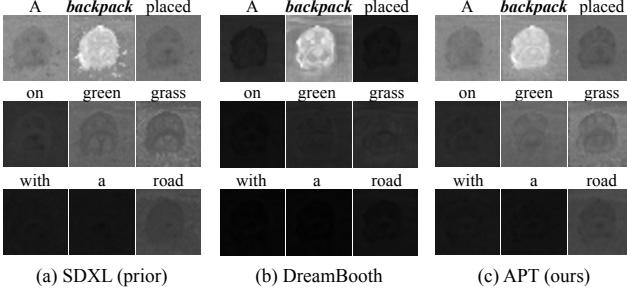
### 3.3. Attention Alignment for Prior Preservation

Overfitting can lead to the loss of prior knowledge specified by the text embeddings, causing the model to generate images that do not accurately reflect the desired context. For example, when learning a concept like a bag, the pre-trained model might generate images that include prior knowledge associations (e.g., a person carrying the bag) even without explicit prompts. In contrast, the fine-tuned model may lose this capability, leading to incoherent images.

To address these issues, we introduce attention alignment for prior knowledge preservation, a regularization technique to align the cross-attention maps of the fine-tuned model with those of the pre-trained model. Let  $\mathbf{A}_{\theta,i}^{(l)}$  and  $\mathbf{A}_{\phi,i}^{(l)}$  denote the cross-attention maps at layer  $l$  and attention head  $i$  for the fine-tuned model  $\theta$  and the pre-trained model  $\phi$ , respectively. We define the attention regularization loss as:

$$\mathcal{L}_{\text{attn}} = \sum_l^{\text{layers}} \frac{1}{H} \left| \sum_{i=1}^H \mathbf{A}_{\theta,i}^{(l)}(x_t; c^*, t) - \sum_{i=1}^H \mathbf{A}_{\phi,i}^{(l)}(x_t; c, t) \right|_2^2, \quad (6)$$

where  $H$  is the number of attention heads. By differentiating between  $c^*$  and  $c$ , we align the attention maps corresponding to the personalized concept with those of the general class, preserving prior knowledge.



**Figure 3. Cross-Attention Map Comparison.** Visualization of cross-attention maps in text-conditioned image generation for (a) SDXL, (b) DreamBooth, and (c) APT. DreamBooth shows changes not only in the class token’s map but also in overall attention maps, indicating shifts in how the model attends to different tokens after personalization.

By applying this regularization to *all* text tokens, we ensure that the model maintains similar attention distributions across all tokens with the pretrained model. As training progresses, we observe that the influence of not only the identifier token but also other tokens changes, as shown in Figure 3. By regularizing all tokens contributing to the representations, we aim to preserve the model’s ability to understand the textual context, retaining the original semantic relationships and allowing the model to generate images that are coherent and contextually appropriate.

### 3.4. Overall Training Objective

The total training loss consists of the proposed regularization terms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{aptDM}} + \lambda_{\text{dist}}(\mathcal{L}_\mu + \mathcal{L}_\sigma) + \lambda_{\text{attn}}\mathcal{L}_{\text{attn}} \quad (7)$$

where the hyperparameters  $\lambda_{\text{dist}}$  and  $\lambda_{\text{attn}}$  control the strength of the regularization terms.

In summary, our proposed method **APT** addresses the challenges of personalizing diffusion models with limited data by introducing adaptive training adjustments, representation stabilization, and attention alignment. By mitigating overfitting in an adaptive manner and preserving the statistical properties and attention distributions of the pretrained model, we enhance the ability to retain prior knowledge and maintain semantic coherence during fine-tuning.

## 4. Experiments

In this section, we evaluate the effectiveness of our proposed APT in personalizing diffusion models with limited reference data. We compare APT with existing techniques through qualitative and quantitative comparisons, a user study, an ablation study, and an analysis of the overfitting indicator. Detailed ablation results and additional evaluations are provided in the Supplementary Material B.

### 4.1. Experimental Setup

We adopt the pretrained Stable Diffusion XL model [22]<sup>2</sup> as the foundation for all experiments. Our evaluations are conducted on commonly used datasets in personalization studies, specifically the **DreamBooth Dataset** [27] and the **Textual Inversion Dataset** [8]. To generate captions for these images, we employ GPT-4o [1], ensuring that the captions emphasize background descriptions while omitting explicit mentions of the target concept. This strategy prevents interference with the learning of the identifier and enables the model to focus on contextual details.

**Baselines** We compare our APT method with the following baseline personalization techniques:

- **DreamBooth:** Combines DreamBooth [27] and Textual Inversion [8] methods for concept learning.
- **Custom Diffusion** [16]: Performs efficient personalization by updating only the key and value of cross-attention.
- **Direct Consistency Optimization (DCO)** [17]: Addresses overfitting by regularizing the denoising process.

**Implementation Details** Most methods, including ours, employ rank-32 LoRA [14] for both U-Net and text encoder with a learning rate of  $5 \times 10^{-5}$  and  $5 \times 10^{-6}$  respectively, using a batch size of 1. Custom Diffusion [16] does not use LoRA but instead fine-tunes the key and value of cross-attention, with a learning rate of  $1 \times 10^{-5}$  and the same batch size of 1. The regularization weights are set to  $\lambda_{\text{dist}} = 30$  and  $\lambda_{\text{attn}} = 3 \times 10^{-4}$ , and the maximum augmentation probability  $p_{\text{max}}$  is 0.8. Further implementation details are provided in Supplementary Material A.

### 4.2. Qualitative Analysis

Figure 4 shows a qualitative comparison between APT and baseline methods using identical text prompts. Our observations are summarized as follows:

- **Scene Context and Background Preservation:** APT generates coherent backgrounds and naturally places objects (e.g., placing a backpack in a landscape), whereas baseline methods often generate overly zoomed-in views.
- **Prior Knowledge Preservation:** Unlike baselines that generate only the object, APT leverages the pretrained model’s prior knowledge to incorporate contextual elements such as human subject.
- **Textural and Stylistic Consistency:** APT replicates textures and styles from the pretrained model while maintaining semantic coherence.
- **Text Alignment:** APT faithfully follows textual instructions, achieving superior alignment with prompt details.

Overall, these qualitative results confirm that APT effectively achieves a balance between preserving contextual

<sup>2</sup>Additional experiments using Stable Diffusion v2.1 are provided in Supplementary Material Section B.4.

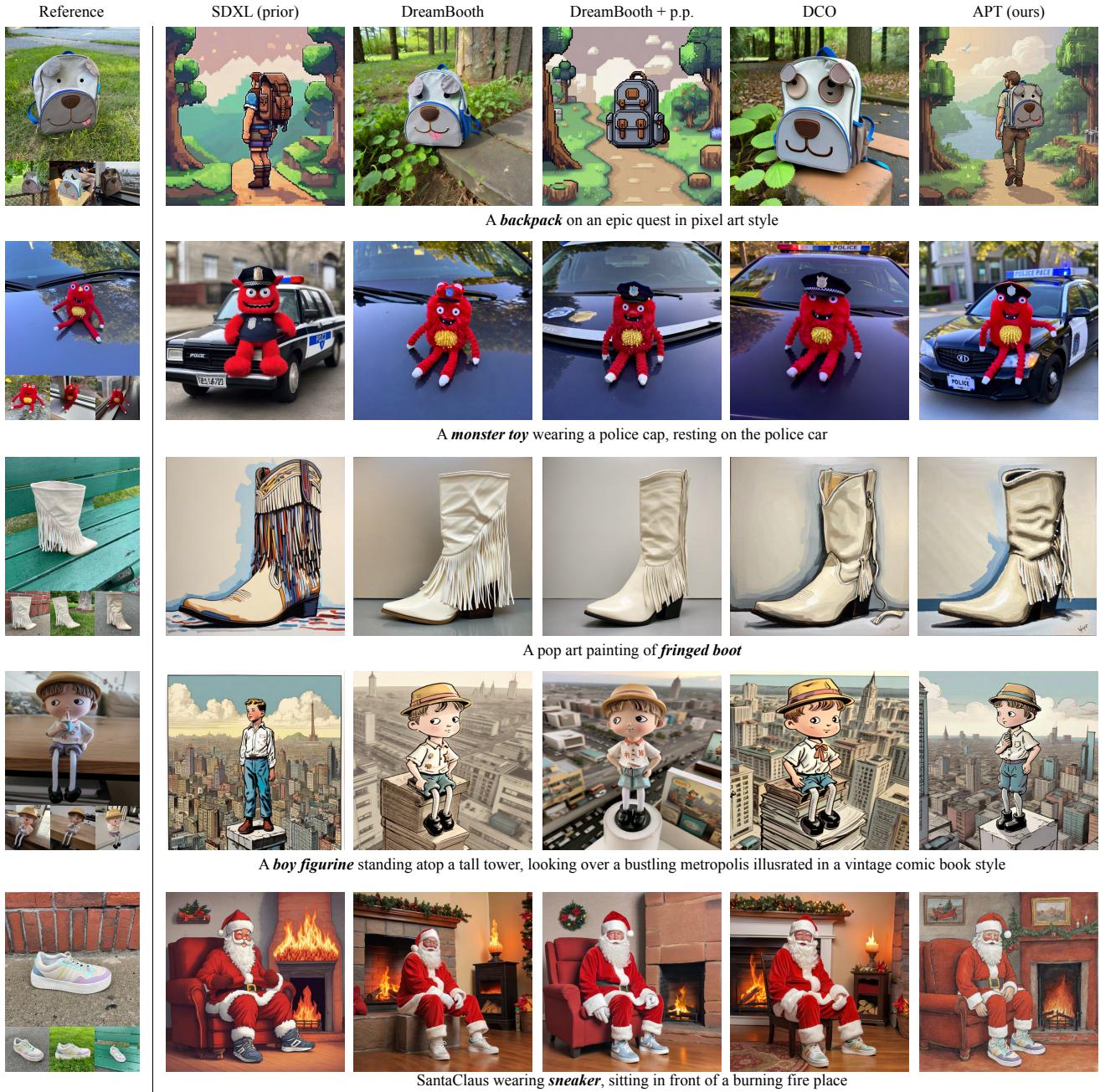


Figure 4. **Qualitative Comparison.** We present images generated by the pretrained model, DreamBooth, DreamBooth with Prior Preservation, DCO, and our method (**APT**) across various data types and styles. Baseline methods tend to memorize textures and generate object-centric images, often lacking prior knowledge such as generating a person without explicit prompts. Objects are frequently zoomed in, with limited contextual and background details. In contrast, our method effectively integrates prior knowledge and generates images with better contextual alignment.

integrity and generating high-fidelity objects. Additional qualitative comparisons can be found in Supplementary material B.1.

### 4.3. Quantitative Analysis

To quantitatively assess performance, we conduct a comprehensive evaluation measuring text-image similarity, image similarity, fidelity, and diversity across different methods

Method	T-I Sim.		I Sim.	Fidelity&Diversity			User Study (%)
	CLIP-T↑	HPSv2↑	DINOv2↑	FID↓	Precision↑	Recall↑	
SDXL (prior)	<b>0.666</b>	<b>0.295</b>	0.625	–	–	–	–
Custom Diffusion [16]	0.662	0.273	0.666	45.530	0.590	0.649	–
DCO [17]	0.662	0.277	<b>0.687</b>	52.298	0.548	0.660	21.1
Base (Dreambooth) [27]	0.661	0.272	<u>0.681</u>	53.130	0.565	0.608	<u>22.8</u>
+ATA	<u>0.664</u>	0.275	0.670	46.872	0.635	0.680	–
+RS	<u>0.664</u>	<u>0.290</u>	0.657	<u>42.663</u>	<b>0.701</b>	<u>0.727</u>	–
+AA (full APT)	<u>0.664</u>	0.288	0.660	<b>41.967</b>	<u>0.669</u>	<b>0.738</b>	<b>56.1</b>

Table 1. Quantitative comparison with baseline methods and ablation study of APT components. For evaluation, we use multiple metrics: Text-Image Similarity measured by CLIP-T and HPSv2 (higher values indicate better text alignment); Image Similarity measured by DINOv2 image-feature similarity (higher values indicate a closer resemblance to reference images); Fidelity&Diversity measured by FID (lower is better) and Precision/Recall (higher is better); and User Study showing the percentage of participants selecting each method based on the criteria: preservation of prior knowledge, ability to capture the identity of reference images, and alignment with the prompt.

using a diverse set of prompts from MS COCO [5] captions. The results can be found in Table 1.

- **Text-Image Similarity:** We use the CLIP-T [24] score and HPSv2 [35] to evaluate how well generated images align with their corresponding text prompts. The CLIP-T score measures the cosine similarity between image and text embeddings, while HPSv2 assesses human preference for image-text alignment. Our method achieves superior text alignment on both metrics, particularly in HPSv2 scores, indicating better adherence to textual prompts while maintaining image quality.
- **Image Similarity:** We compute the image similarity using DINOv2 features [20], which capture the semantic information of images. The similarities are calculated as the average pairwise cosine similarity between generated and reference images. Our method effectively preserves subject identity, outperforming SDXL while being comparable to or slightly lower than other baselines. This is due to our stronger emphasis on scene context over object-centric generation, which results in reduced zoomed-in artifacts in the generated images. Since DINOv2 similarity scores tend to favor closely cropped, object-centric images, our method’s slightly lower similarity score reflects its ability to incorporate broader scene context rather than a deficiency in concept capture.
- **Fidelity & Diversity:** While precision and recall traditionally measure the fidelity and diversity of generated samples with respect to the real data distribution [30], applying these metrics directly to diffusion model personalization is challenging. The few-shot nature of reference images prevents a reliable estimation of the real data distribution. Instead, we evaluate our method from a prior preservation perspective, measuring how well the personalized model maintains SDXL’s generation capabilities. We establish two datasets: a source dataset generated us-

ing SDXL and a target dataset generated using personalized models. We then measure FID, Precision, and Recall between these source and target datasets. Our method outperforms other approaches, effectively preserving the original generation capabilities of SDXL while inheriting both its fidelity and diversity characteristics.

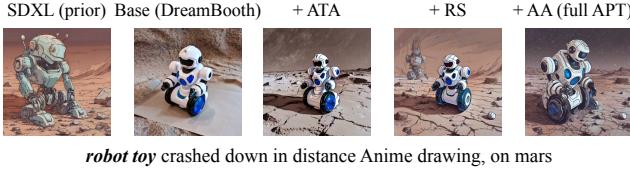
Overall, our method achieves competitive quantitative performance, validating its effectiveness in personalizing diffusion models with limited data.

#### 4.4. User Study

We also conduct a user study to assess how well different models achieve personalization from a human alignment perspective. For simplicity in evaluation, we select DreamBooth and DCO as major baselines for comparison with our method. 20 participants blindly evaluate images from all three methods across 20 different prompts, with reference images and SDXL-generated images provided as prior knowledge on the following criteria (refer to Supplementary Material C for details):

- Assess the **text alignment** between the text prompt and the generated image, selecting the image that best reflects the detailed features of the text prompt.
- Evaluate the **identity similarity** between objects in the training data and those in the generated images, along with the overall **image quality**.
- Compare with images generated by the pretrained model, considering whether the generated images effectively **preserve prior knowledge** and are contextually appropriate.

As shown in Table 1, 56.1% of the participants preferred the images generated by APT, compared to 22.8% and 21.1% for DreamBooth and DCO, respectively. This indicates that our method better aligns with the prompt and generates more visually appealing images than comparison methods.



**Figure 5. Ablation Study of APT Components.** We evaluate the contribution of each component in our method by incrementally adding Adaptive Training Adjustment (ATA), Representation Stabilization (RS), and Attention Alignment (AA) to Base (DreamBooth).

#### 4.5. Ablation Study

To assess the contribution of each component in our method, we perform an ablation study by incrementally adding each component and observing the qualitative and quantitative effects (see Figure 5 and Table 1).

1. **Base (DreamBooth):** Base method without any of our proposed components
2. **+ATA:** Base + Adaptive Training Adjustment. With +ATA, zoomed-in artifacts are significantly reduced while improving Precision, Recall, and FID compared to Base, demonstrating the effectiveness of adaptive training strategies in mitigating overfitting.
3. **+RS:** Base + ATA + Representation Stabilization. Adding +RS further reduces texture memorization effects by reducing distribution shifts, significantly improving HPSv2 and the model’s generalization ability while preserving the text fidelity of SDXL.
4. **+AA (full APT):** Base + ATA + RS + Attention Alignment. With the full APT method (+AA), we observe better preservation of prior knowledge and improved metrics through regularization, allowing the personalized subject to be generated in a more coherent and contextually appropriate manner.

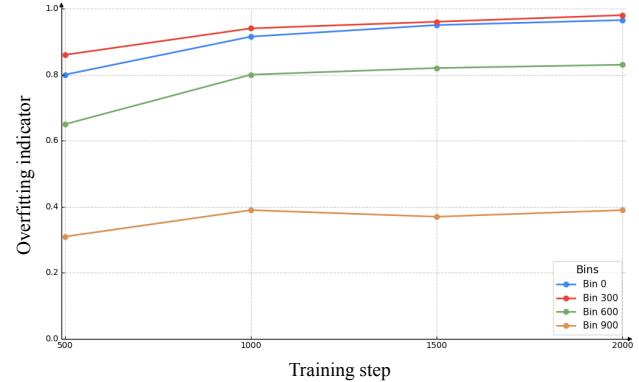
As we add each component, we observe progressive improvements in image quality, text alignment, and preservation of prior knowledge. The full APT method generates the most coherent and contextually appropriate images. Additional examples and details are provided in Supplementary Material for further reference (see B.2).

#### 4.6. Analysis of Overfitting Indicator

We analyze the behavior of the overfitting indicator  $\gamma_t$  over training steps and time step bins to understand its influence on adaptive training adjustments. Figure 6 presents a plot of  $\gamma_t$  across training iterations for different bins.

We observe that the overfitting indicator  $\gamma_t$  increases more significantly for later time step bins (low noise levels) than for early timesteps (high noise levels). This indicates that overfitting occurs more rapidly at steps closer to the final image reconstruction, where the model begins to memorize specific details of the training data. The adaptive data

augmentation and loss weighting respond accordingly, adjusting the training dynamics to mitigate overfitting where it is most pronounced. This adaptive mechanism helps maintain the stability of the denoising trajectory and preserves prior knowledge by dynamically adjusting to varying overfitting tendencies across different time steps.



**Figure 6. Overfitting Indicator** The overfitting indicator  $\gamma_t$  is plotted across training iterations for different time step bins.

#### 5. Limitations

While our proposed APT method effectively mitigates overfitting and preserves prior knowledge, it has certain limitations. The trade-off between preserving prior knowledge and learning new concepts is a fundamental challenge in text-to-image personalization. Although our work approaches Pareto-optimal solutions, some challenges remain. For example, when personalizing a “monster toy” intended to be cute, the strong prior associated with the word “monster” may cause the model to generate images with more monstrous appearances than desired. This issue arises because the identifier used during personalization is heavily influenced by the class word chosen for initialization. Adjusting the regularization weight  $\lambda_{\text{attn}}$  associated with attention alignment can alleviate this problem by allowing more flexibility in how the model integrates prior knowledge. However, this introduces sensitivity to hyperparameters, which remains a limitation as it requires careful tuning for different concepts.

#### 6. Conclusion

We have presented APT, a novel method for personalizing diffusion models with limited data. By incorporating adaptive training adjustments, representation stabilization, and attention alignment, APT effectively mitigates overfitting and preserves prior knowledge. Our experiments demonstrate that APT outperforms existing methods, providing a robust solution for personalized generative modeling. Further research directions are discussed in Supplementary Material D.

## Acknowledgement

This research was supported by Seoul National University of Science and Technology (2024-0200).

## 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. [5](#) [1](#)
- [2] Yuval Alaluf, Elad Richardson, Gal Metzer, and Daniel Cohen-Or. A neural space-time representation for text-to-image personalization. *ACM Transactions on Graphics (TOG)*, 42(6):1–10, 2023. [3](#) [7](#)
- [3] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–12, 2023. [2](#)
- [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. *Computer Science*. <https://cdn.openai.com/papers/dall-e-3.pdf>, 2(3):8, 2023. [2](#)
- [5] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. [7](#)
- [6] Hyungjin Chung, Jeongsol Kim, Geon Yeong Park, Hyelin Nam, and Jong Chul Ye. Cfg++: Manifold-constrained classifier free guidance for diffusion models. *arXiv preprint arXiv:2406.08070*, 2024. [1](#)
- [7] Hyungjin Chung, Jeongsol Kim, Geon Yeong Park, Hyelin Nam, and Jong Chul Ye. Cfg++: Manifold-constrained classifier free guidance for diffusion models. *arXiv preprint arXiv:2406.08070*, 2024. [2](#)
- [8] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2022. [2](#) [3](#) [5](#) [7](#)
- [9] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7323–7334, 2023. [2](#) [3](#)
- [10] Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, and Baining Guo. Efficient diffusion training via min-snr weighting strategy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7441–7451, 2023. [2](#)
- [11] Shaozhe Hao, Kai Han, Shihao Zhao, and Kwan-Yee K Wong. Vico: Plug-and-play visual condition for personalized text-to-image generation. *arXiv preprint arXiv:2306.00971*, 2023. [3](#) [7](#)
- [12] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. [2](#) [1](#)
- [13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. [2](#)
- [14] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. [3](#) [5](#) [1](#)
- [15] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020. [3](#)
- [16] 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*, pages 1931–1941, 2023. [3](#) [5](#) [7](#)
- [17] Kyungmin Lee, Sangkyung Kwak, Kihyuk Sohn, and Jinwoo Shin. Direct consistency optimization for compositional text-to-image personalization. *NeurIPS*, 2024. [2](#) [3](#) [5](#) [7](#) [1](#) [6](#)
- [18] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. [1](#)
- [19] 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. [2](#)
- [20] Maxime Oquab, Timothée Darct, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. [7](#)
- [21] Lianyu Pang, Jian Yin, Baoquan Zhao, et al. Attndream-booth: Towards text-aligned personalized text-to-image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. [2](#) [3](#) [7](#)
- [22] Dustin Podell, Zion English, Kyle Lacey, et al. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024. [5](#) [1](#) [2](#)
- [23] Zeju Qiu, Weiyang Liu, Haiwen Feng, Yuxuan Xue, Yao Feng, Zhen Liu, Dan Zhang, Adrian Weller, and Bernhard Schölkopf. Controlling text-to-image diffusion by orthogonal finetuning. *Advances in Neural Information Processing Systems*, 36:79320–79362, 2023. [3](#) [7](#)
- [24] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. [7](#)
- [25] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. [2](#)
- [26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. [2](#)

- [27] 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*, pages 22500–22510, 2023. [2](#), [3](#), [5](#), [7](#), [6](#)
- [28] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10, 2022. [2](#)
- [29] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamvar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Raphael Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 2022. [2](#)
- [30] Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. *Advances in neural information processing systems*, 31, 2018. [7](#)
- [31] Dazhong Shen, Guanglu Song, Zeyue Xue, Fu-Yun Wang, and Yu Liu. Rethinking the spatial inconsistency in classifier-free diffusion guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9370–9379, 2024. [2](#)
- [32] Kaede Shiohara and Toshihiko Yamasaki. Face2diffusion for fast and editable face personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6850–6859, 2024. [2](#)
- [33] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2020. [2](#)
- [34] 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. [3](#)
- [35] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023. [7](#)
- [36] 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. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022. [2](#)
- [37] Yu Zeng, Vishal M Patel, Haochen Wang, Xun Huang, Ting-Chun Wang, Ming-Yu Liu, and Yogesh Balaji. Jedi: Joint-image diffusion models for finetuning-free personalized text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6786–6795, 2024. [2](#)
- [38] 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*, pages 4764–4774, 2024. [2](#)
- [39] Zhengli Zhao, Sameer Singh, Honglak Lee, Zizhao Zhang, Augustus Odena, and Han Zhang. Improved consistency regularization for gans. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11033–11041, 2021. [3](#)

# APT: Adaptive Personalized Training for Diffusion Models with Limited Data

## Supplementary Material

### A. Implementation Details

**Additional Details** All experiments are conducted using a single NVIDIA A100 GPU. For Representation Stabilization, we utilize the hidden states from the Upblocks of the U-Net at resolutions of  $32 \times 32$  and  $64 \times 64$ . Additionally, for Attention Alignment, we employ the attention maps from the same Upblocks. We use the AdamW optimizer [18] for training all models. The learning rate and other optimizer hyperparameters are set as described in the main text. In Adaptive Data Augmentation, we apply zoom-out transformations with scales ranging from 1 to 3 and rotations within  $\pm 15$  degrees. We acknowledge that further experiments with additional augmentation types could be beneficial and are left for future work. For the Exponential Moving Average (EMA) calculations, we set the smoothing factor  $\alpha$  to 0.1. All generated images are generated using a Classifier-Free Guidance (CFG) [12] with scale of 7.5. For DCO [17], to ensure a fair comparison, we use only CFG without Reward Guidance.

**GPT-4o Caption Details** Building upon Comprehensive Caption [17], we employ GPT-4o [1] to generate captions that emphasize on the background and context rather than the primary concept, allowing the token to learn the concept as directly as possible. We provide the reference data into GPT-4o and instruct it to describe each image, focusing on the surroundings and context while keeping the description of the central object as simple as possible. We observe that when prompts contain detailed descriptions of the concept, the model struggles to learn those details effectively. By shifting the focus of captions to background and contextual elements, we ensure that the model learns rich and diverse information. This approach not only enhances the learning of the desired concept through the token but also prevents the model from learning about non-target objects. By omitting detailed descriptions of the concept’s color, texture, and other fine-grained details, we promote more robust learning and achieve better generalization when generating images conditioned on the learned concept.

**Computations** Our method requires an extra forward pass to retrieve the intermediate features of SDXL [22], which increases computational overhead—an approach also employed by the state-of-the-art method, DCO [17]. However, since LoRA [14] loaded into SDXL can be toggled on or off during the forward pass, our approach requires only the additional memory needed for the intermediate features, without the need to load a separate pretrained model.

### B. Additional Experimental Results

#### B.1. Qualitative Comparisons

In Figure 9 and 10, we present additional qualitative comparisons between APT and baseline methods across diverse datasets and text prompts to demonstrate our model’s superior performance. Our qualitative analysis reveals several key advantages of APT over existing approaches in four critical aspects described in Section 4.2. The baseline methods exhibit notable limitations in maintaining scene context and integrating prior knowledge, often generating overly focused, decontextualized images. For instance, when generating images of sneakers, baseline methods tend to generate isolated views that fail to capture the impressionist style specified in the prompt, while APT successfully incorporates these objects into coherent, prompt-aligned scenes that reflect the artistic direction.

APT demonstrates remarkable capability in preserving prior knowledge from pretrained models, particularly in scenarios involving artistic style integration. When generating images of an alarm clock, APT successfully captures both the Magritte-style surrealist background and the distinctive texture of LEGO building blocks, while baseline methods struggle to maintain these artistic elements, often defaulting to conventional representations that lack the specified stylistic characteristics. This showcases the ability of APT to simultaneously handle multiple style requirements while maintaining object consistency.

#### B.2. Ablation Study

We provide additional ablation results and analysis (see Table 1 and Figure 7) to further demonstrate the impact of each component in our proposed APT framework. These results complement Section 4.5 and offer deeper insights into how each component contributes to mitigating overfitting and preserving prior knowledge.

**Adaptive Training Adjustment (ATA)** ATA immediately improves the baseline by mitigating overfitting. As shown in Table 1, applying ATA to the base model results in a modest increase in text-image similarity scores (with slight improvements in both CLIP-T and HPSv2) and a significant reduction in FID, which indicates better fidelity and diversity. Qualitatively, as illustrated in Figure 7 (3<sup>rd</sup> column), the “zoomed-in” effect observed in the base model’s outputs is eliminated with ATA. The personalized object is no longer unnaturally enlarged or forced into the center; instead, it is rendered with greater flexibility in layout. This

demonstrates that by introducing adaptive data augmentation and loss weighting, ATA effectively prevents the model from overfitting to a specific region or scale, thereby allowing for more natural object placement and pose variation.

**Representation Stabilization (RS)** Building on ATA, the addition of RS further improves the model’s performance. In Table 1, RS improves metrics related to prior preservation and alignment—for instance, increasing HPSv2 (indicating better prompt alignment) while slightly decreasing DINOv2 similarity (suggesting reduced over-tuning to reference details). Figure 7 (4<sup>th</sup> column) confirms that RS stabilizes intermediate representations during fine-tuning, which reduces the over-saturation of the subject’s texture. By adjusting the distribution of latent features, RS prevents direct texture memorization, enabling the model to generalize better across different scenes and lighting conditions, while preserving the pretrained knowledge to adhere to the text prompt structure.

**Attention Alignment (AA)** Finally, incorporating AA (yielding the full APT model) unifies the benefits of the previous components and further refines the output. As shown in Table 1, AA helps the model maintain high text-image similarity while achieving low FID values. Supplementary metrics such as Recall also improve with AA, indicating enhanced output diversity. Figure 7 (5<sup>th</sup> column) demonstrates that AA improves semantic coherence: when applied, a personalized figurine is generated not only with its identity preserved but also with background elements and contextual cues that closely align with the prompt. AA achieves this by explicitly aligning the model’s attention maps with those of the pretrained model, ensuring that attention is distributed across all prompt elements rather than being overly concentrated on the new concept token.

**Overall Analysis** The supplementary ablation study confirms that each component in APT contributes both individually and synergistically. ATA primarily mitigates spatial overfitting by freeing the object from a constrained, zoomed-in view. RS addresses feature-space overfitting by maintaining generalizable intermediate representations, and AA combats attention overfitting by ensuring a balanced focus across the entire prompt and scene. Although minor trade-offs (such as a slight decrease in precision with AA) are observed, they are more than compensated for by major gains in diversity and overall image coherence. Together, these results reinforce our claim that APT’s components are complementary and collectively enable state-of-the-art performance in personalized diffusion model training with limited data.

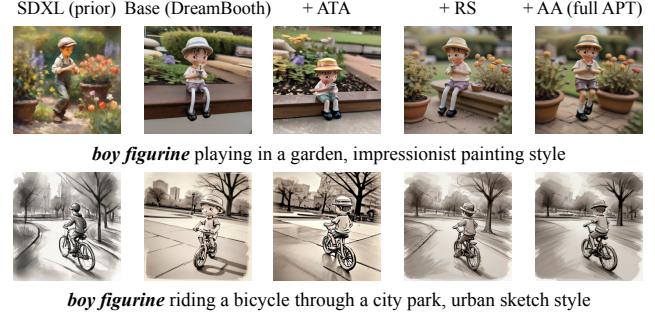


Figure 7. **Additional Ablation Study of APT Components.** We evaluate the contribution of each component in our method by incrementally adding Adaptive Training Adjustment (ATA), Representation Stabilization (RS), and Attention Alignment (AA) to Base (DreamBooth).

### B.3. Motivation for Adaptive Loss Weighting

Given a paired dataset of images  $\mathbf{x}$  and captions  $\mathbf{c}$ , diffusion models are trained using a simplified version of the variational bound objective [13, 26]:

$$\mathcal{L}_{\text{simple}}(\theta; \mathcal{D}) := \mathbb{E}_{(\mathbf{x}, \mathbf{c}) \sim \mathcal{D}, \epsilon, t} [\omega(t) \|\epsilon - \epsilon_\theta(\mathbf{x}_t; \mathbf{c}, t)\|^2], \quad (8)$$

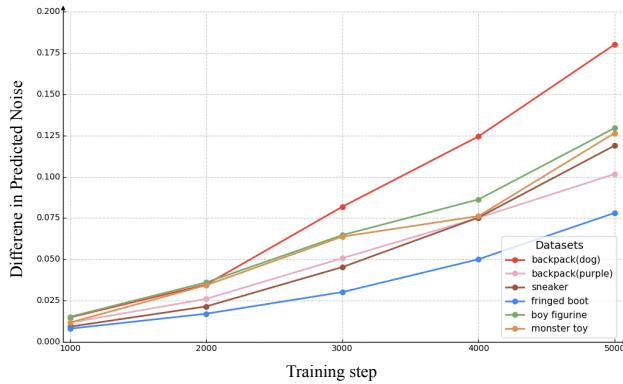
where  $\mathbf{x}_t = \alpha_t \mathbf{x}_{t-1} + \sigma_t \epsilon$  for  $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ ,  $t \sim \mathcal{U}(0, T)$ .  $\omega(t)$  is a weighting function allowing the model to focus on more challenging denoising tasks at larger timestep  $t$  and make better sample quality. Min-SNR [10] improves the convergence speed of training by considering the reverse process as a multi-task problem with varying difficulty levels and applying different clamped loss weights for each timestep interval.

However, since the training dynamics of personalizing diffusion models with limited data vary across different datasets, this necessitates excessive time and effort for hyperparameter optimization. Figure 8 illustrates the differences between the predicted noise of the pretrained SDXL model [22] and that of the model fine-tuned using the DreamBooth [27] method, as follows:

$$\Delta \text{Noise} = \|\epsilon_\phi(\mathbf{x}_t; \mathbf{c}, t) - \epsilon_\theta(\mathbf{x}_t; \mathbf{c}, t)\|^2 \quad (9)$$

As training progresses, the model loses the original distribution due to excessive shifts in the noise prediction, focusing solely on memorizing the training data and consequently degrading the model’s ability to generalize to unseen prompts. This phenomenon appears similar across all datasets, but different overfitting patterns can be observed. At the end of training, the predicted noise difference between the model trained on the backpack (dog) dataset and the pretrained model is more than twice as large as that of the model trained on the fringed boot dataset. While severe overfitting may occur in specific datasets, this pattern does not generalize across all objects. Against this background,

in Section 3.1, we introduce an Adaptive Overfitting Indicator that quantitatively measures the degree of overfitting during training in a dataset-dependent manner. Since the degree of overfitting varies across different datasets, our indicator adjusts adaptively during training. Additionally, we design a weighting scheme to reduce the impact of the loss accordingly when overfitting is detected, allowing the weights to vary based on the dataset rather than remaining fixed, as in previous approaches.



**Figure 8. Difference in Predicted Noise.** The difference in predicted noise between SDXL (prior) and DreamBooth [27] models is plotted over training iterations. Since the degree of overfitting varies across different datasets, we were motivated to detect overfitting during training and adjust the impact of the loss accordingly.

#### B.4. Application to Stable Diffusion V2.1

To demonstrate that our proposed APT is not only applicable to Stable Diffusion XL (SDXL) but also competitive when applied to other models, we conduct experiments using Stable Diffusion V2.1. Most existing personalization methods have been developed and evaluated on Stable Diffusion versions 1.4 or 2.1; thus, experimenting with V2.1 allows for a broader comparison with these methods.

In Figure 11, we compare APT with other methods based on Stable Diffusion V2.1, including DreamBooth [8, 27], NeTI [2], ViCo [11], OFT [23], and AttnDreamBooth [21]. All images except those generated by our method are directly taken from AttnDreamBooth [21].

For Stable Diffusion V2.1, we observe that the convergence speed of the overfitting indicator  $\gamma$  differed from that in SDXL. Specifically,  $\gamma$  converges more rapidly due to the characteristics of the model. To account for this, we adjust the calculation of  $\gamma$  by using  $T/10$  instead of  $T$  in the exponential function, where  $T$  is the total number of diffusion steps. All other hyperparameters are kept the same as in our experiments with SDXL.

We note that in models like Stable Diffusion V2.1, which have lower generation quality compared to SDXL, preserving prior knowledge can sometimes negatively affect the generated images. This is likely due to the limited capacity

of the model to balance incorporating new concepts while maintaining existing knowledge. Despite this challenge, our method still outperforms the baselines across various styles and contexts by effectively preserving prior knowledge.

### C. User Study

In this section, we provide a detailed explanation of how the user study described in Section 4.4 is conducted. Participants are presented with the following materials:

- **Reference Images:** The original images representing the target concept that the model was trained to learn.
- **Prior Images:** Images generated by the pretrained model (SDXL) using the same noise seed and prompts without any personalization.
- **Prompts:** The text descriptions used to generate images from the models.

Based on these materials, participants are asked to evaluate the generated images by considering the following aspects:

1. **Text Alignment:** Does the generated image align well with the given text prompt?
2. **Identity Preservation:** Is the generated image similar to the reference images?
3. **Prior Similarity:** Is the generated image similar to the composition of the prior image generated by the pretrained model?

Participants are instructed to choose the image that best met all the criteria. Figure 12 shows the interface presented to users during the study. The results of the user study are summarized in Table 1.

### D. Future Work

In this section, we discuss potential areas for improvement and future research directions based on our observations.

#### D.1. Reducing Memory and Computational Overhead

Our method requires forwarding both the pretrained model  $\phi$  and the fine-tuned model  $\theta$  and comparing their attention maps and intermediate representations. This process requires more memory and computations, especially since attention maps from all layers are considered.

To address this issue, future work could focus on optimizing the computation by selecting only a subset of layers or resolutions for attention alignment and representation stabilization. For example, using attention maps and hidden states from specific layers or resolutions (e.g., only higher resolutions) that have the most impact on model performance could reduce computational load without significantly affecting the results.

## D.2. Combining Attention Alignment and Representation Stabilization

Attention alignment and representation stabilization are closely related, as both aim to preserve the model’s internal structures and prior knowledge. Given their close relationship, there is potential to combine these two components into a unified regularization term.

By formulating a joint regularization that considers both the attention maps and the hidden states simultaneously, we may achieve similar or improved performance with reduced computational complexity. Exploring this possibility could lead to a more efficient method that maintains the benefits of both components while mitigating computational overhead.

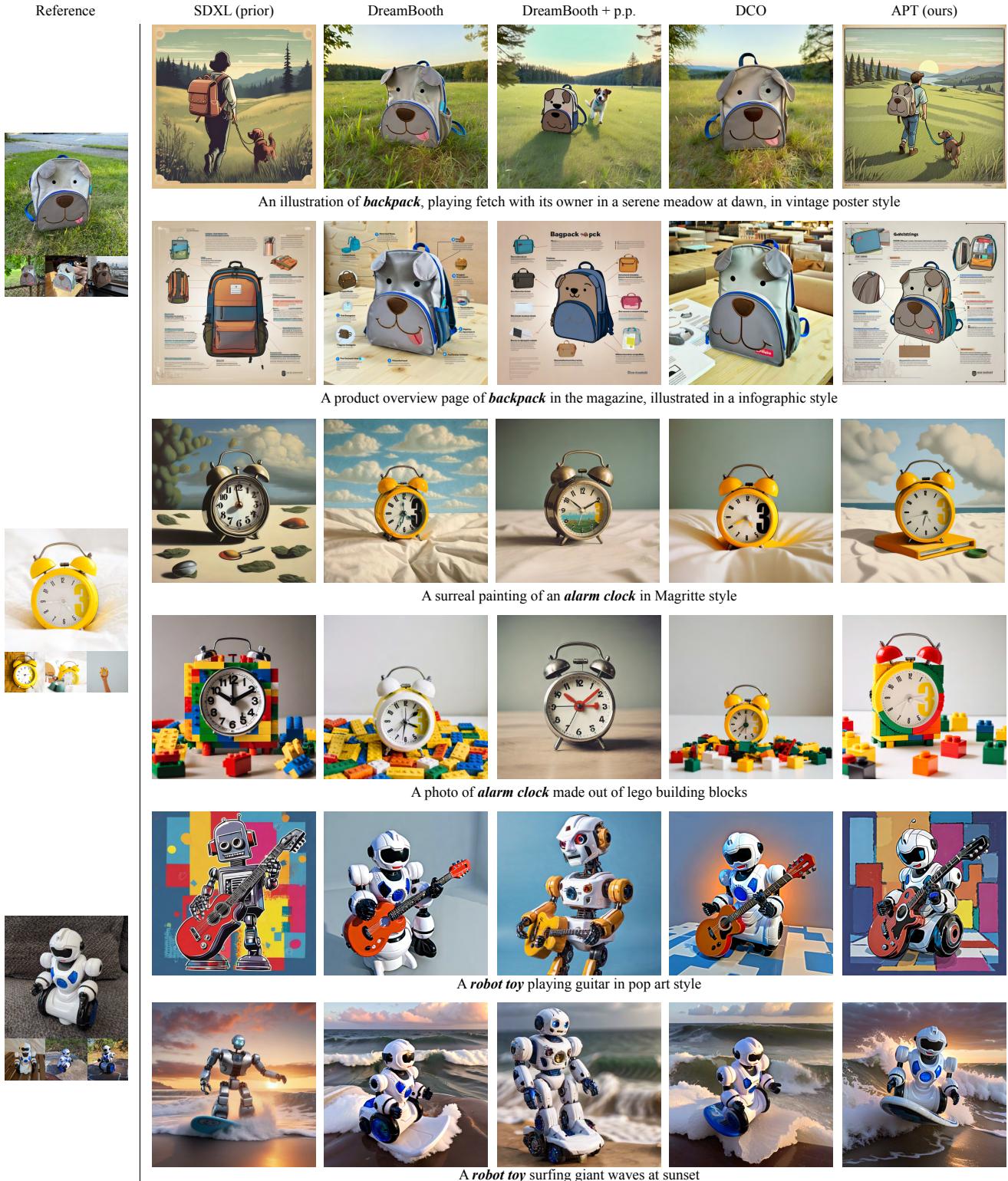


Figure 9. **Additional Qualitative Comparison.** We present four images generated by our method and two images from each of the baseline methods, including SDXL, DreamBooth [27], DreamBooth with prior preservation loss, and DCO [17]. Our method demonstrates superior performance in prior preservation, including text alignment, compared to these baselines.

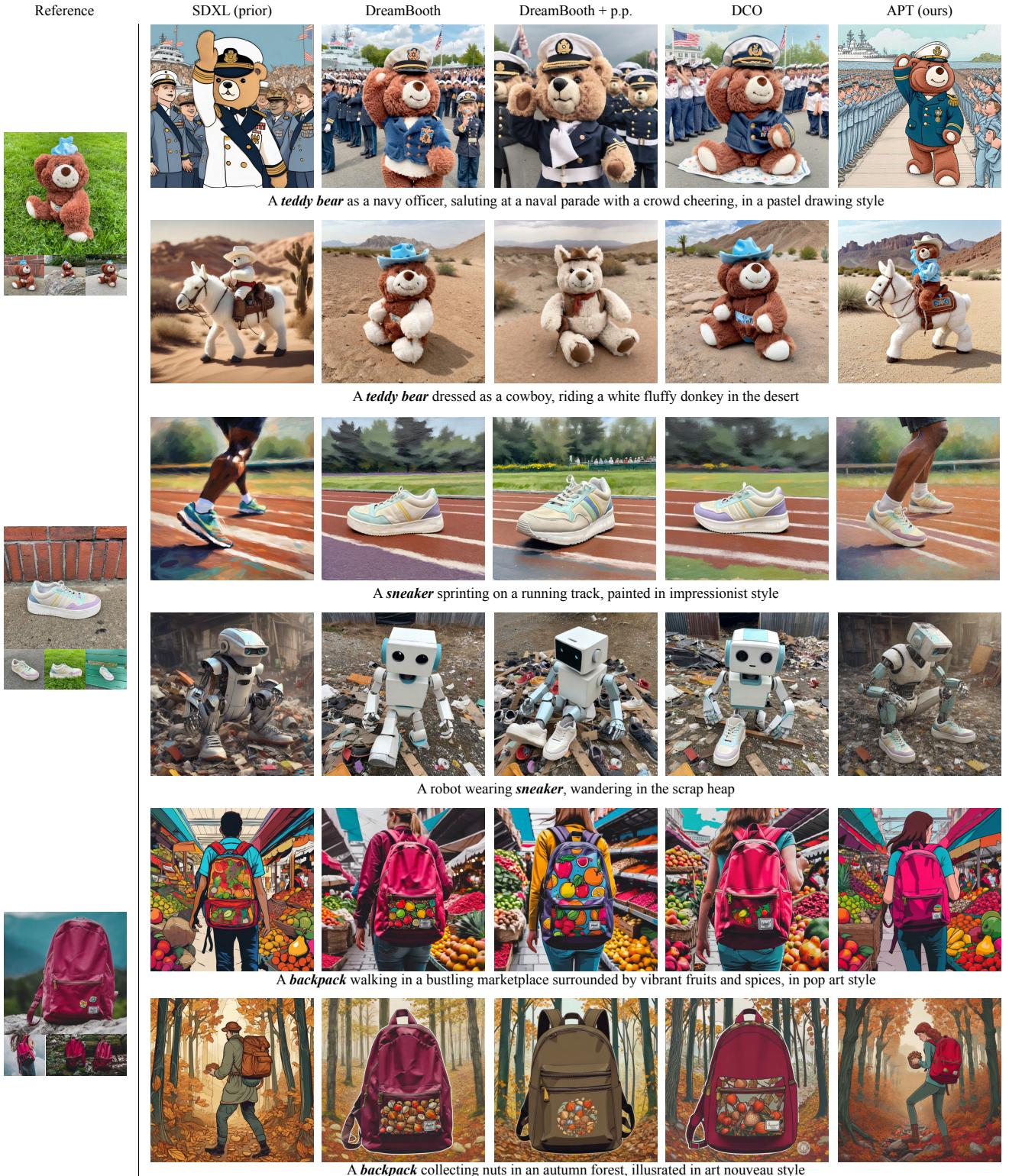
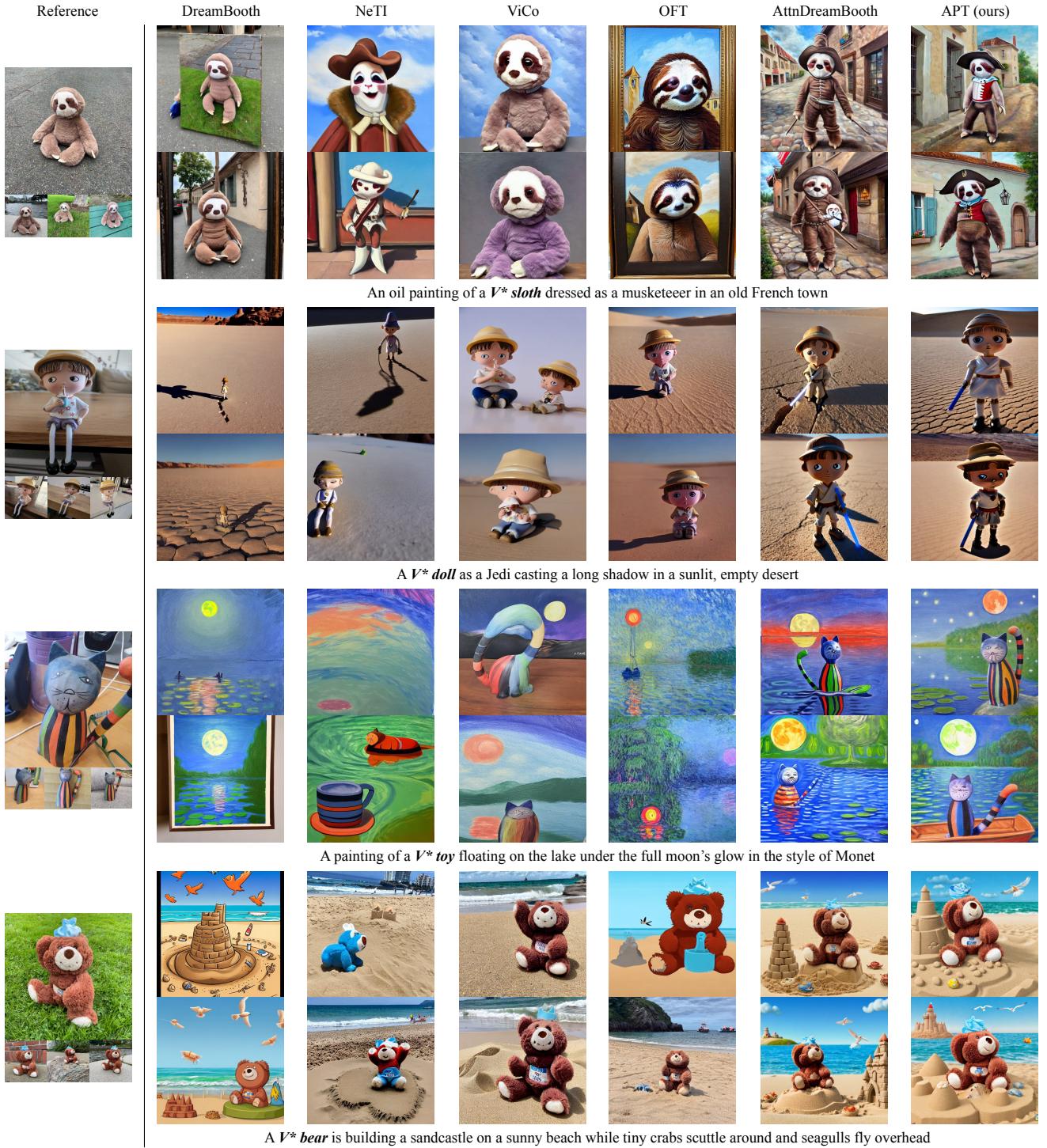


Figure 10. **Additional Qualitative Comparison.** We present four images generated by our method and two images from each of the baseline methods, including SDXL, DreamBooth [27], DreamBooth with prior preservation loss, and DCO [17]. Our method demonstrates superior performance in prior preservation, including text alignment, compared to these baselines.



**Figure 11. Additional Qualitative Comparison on Stable Diffusion V2.1.** We compare APT with other methods which are based on Stable Diffusion V2.1., including DreamBooth [8, 27], NeTI [2], ViCo [11], OFT [23], and AttnDreamBooth [21]. Two images from each of the baseline methods are collected from AttnDreamBooth [21]. Our method outperforms baselines across various styles and contexts by effectively preserving prior knowledge.

Please choose your favorite image among the following three generated images.

When selecting an image, refer to the criteria below:

- Which image aligns well with the given text prompt?
- The top-left image is an example from the training data. which image is more similar to the reference?
- The top-right image is generated by general-purpose image generation model. Which image is more similar to the composition of the prior image?



Reference



SDXL (prior)

Prompt: Oil painting of **backpack** in Seattle during a snowy full moon night



Figure 12. **User Study Example.** This shows the interface presented to users during the study.

