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 32×32 and 64×64. 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 ± 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 α 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

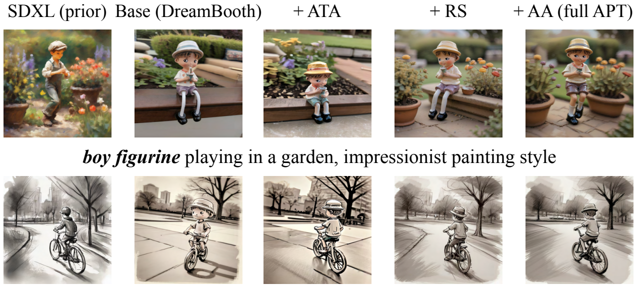
We provide additional ablation results and analysis (see Table 1 and Figure 7) to further demonstrate the impact of each component in our proposed APT framework. These results complement Section 4.5 and offer deeper insights into how each component contributes to mitigating overfitting and preserving 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$^{\text{rd}}$ 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 (4th 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 (5th 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.



boy figurine playing in a garden, impressionist painting style

boy figurine riding a bicycle through a city park, urban sketch style

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]:

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

where $\mathbf{x}_t = \alpha_t\mathbf{x}_{t-1} + \sigma_t\boldsymbol{\epsilon}$ for $\boldsymbol{\epsilon} \sim \mathcal{N}(0,\mathbf{I}), t \sim \mathcal{U}(0,T)$. $\omega(t)$ 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:

$$\Delta Noise = \|\boldsymbol{\epsilon}_\phi(\mathbf{x}_t;\mathbf{c},t) - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t;\mathbf{c},t)\|^2 \tag{9}$$

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.

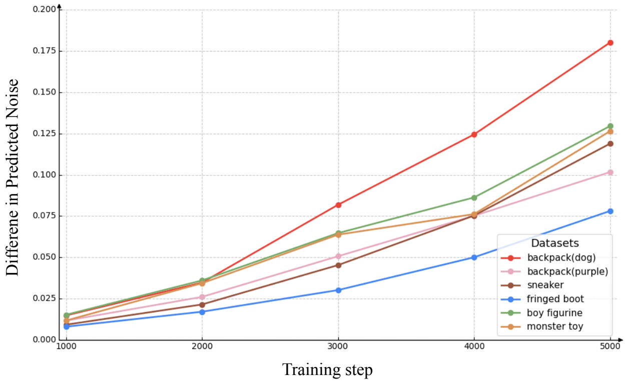


Figure 8. **Difference in Predicted Noise.** The difference in predicted noise between SDXL (prior) and DreamBooth [27] models is plotted over training iterations. Since the degree of overfitting varies across different datasets, we were motivated to detect overfitting during training and adjust the 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 γ differed from that in SDXL. Specifically, γ converges more rapidly due to the characteristics of the model. To account for this, we adjust the calculation of γ 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

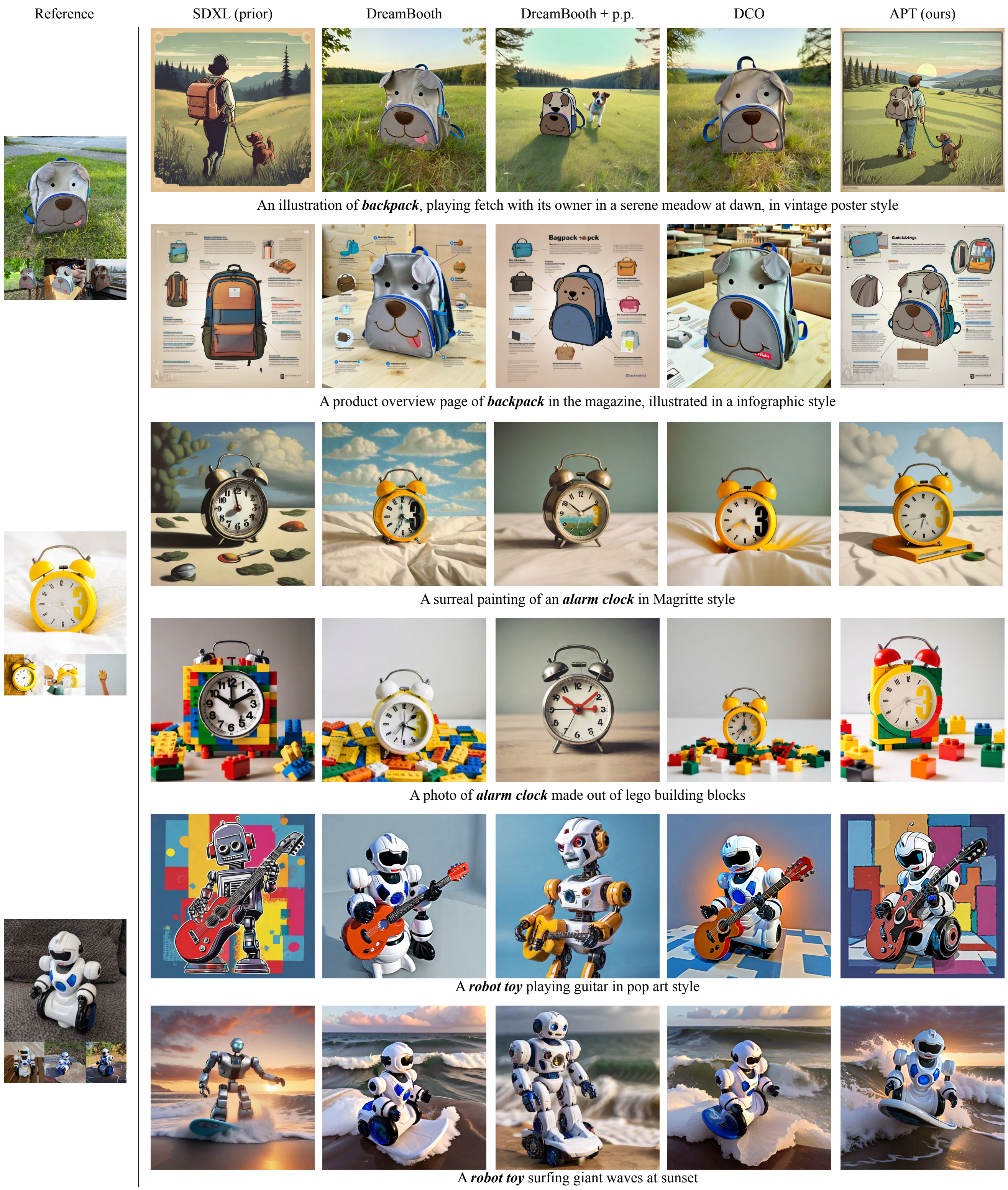
Our method requires forwarding both the pretrained model ϕ and the fine-tuned model θ and comparing their attention maps and intermediate representations. This process requires more memory and computations, especially since attention maps from all layers are considered.

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.



An illustration of ***backpack***, playing fetch with its owner in a serene meadow at dawn, in vintage poster style

A product overview page of ***backpack*** in the magazine, illustrated in a infographic style

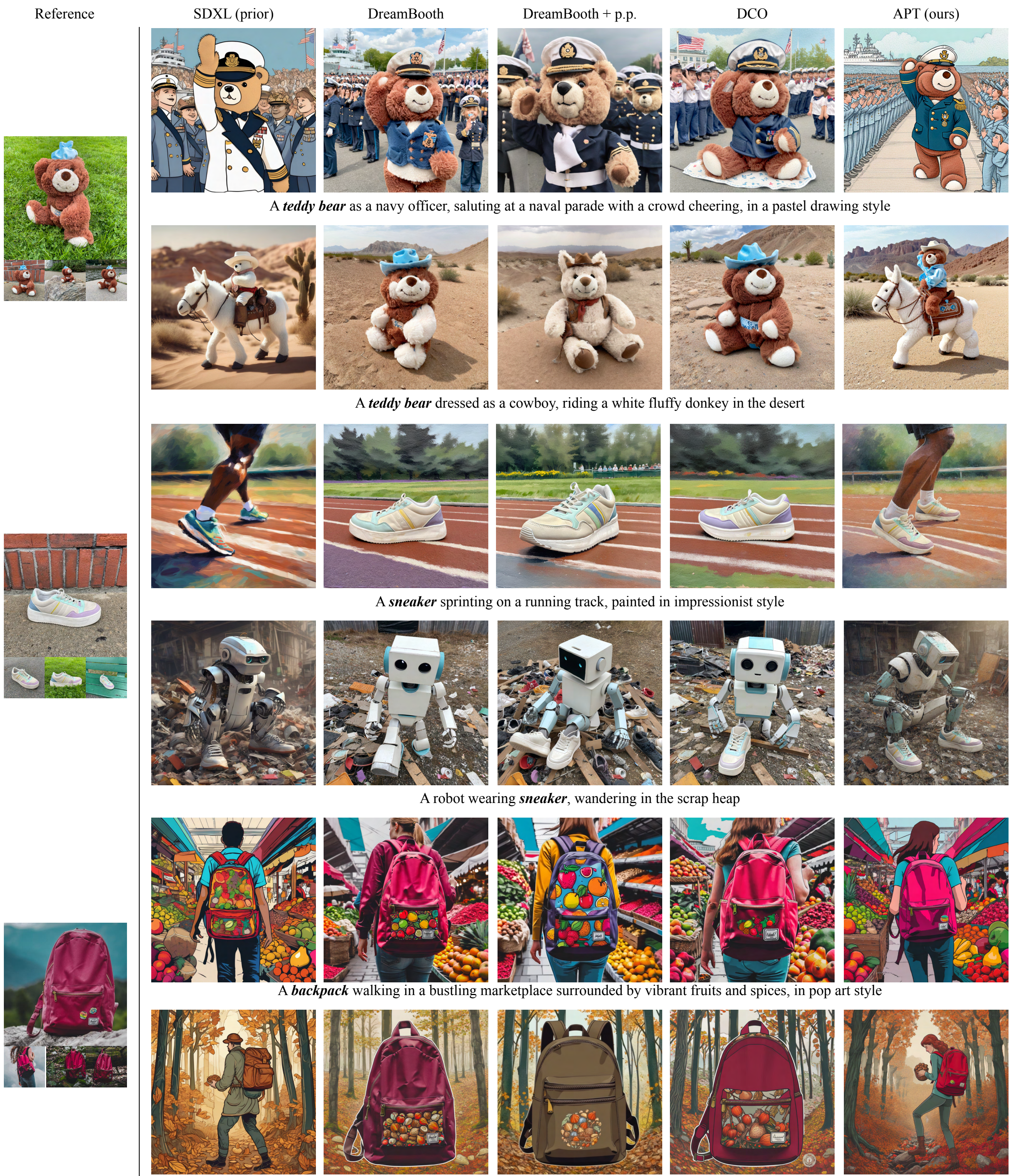
A surreal painting of an ***alarm clock*** in Magritte style

A photo of ***alarm clock*** made out of lego building blocks

A ***robot toy*** playing guitar in pop art style

A ***robot toy*** surfing giant waves at sunset

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.



A ***teddy bear*** as a navy officer, saluting at a naval parade with a crowd cheering, in a pastel drawing style

A ***teddy bear*** dressed as a cowboy, riding a white fluffy donkey in the desert

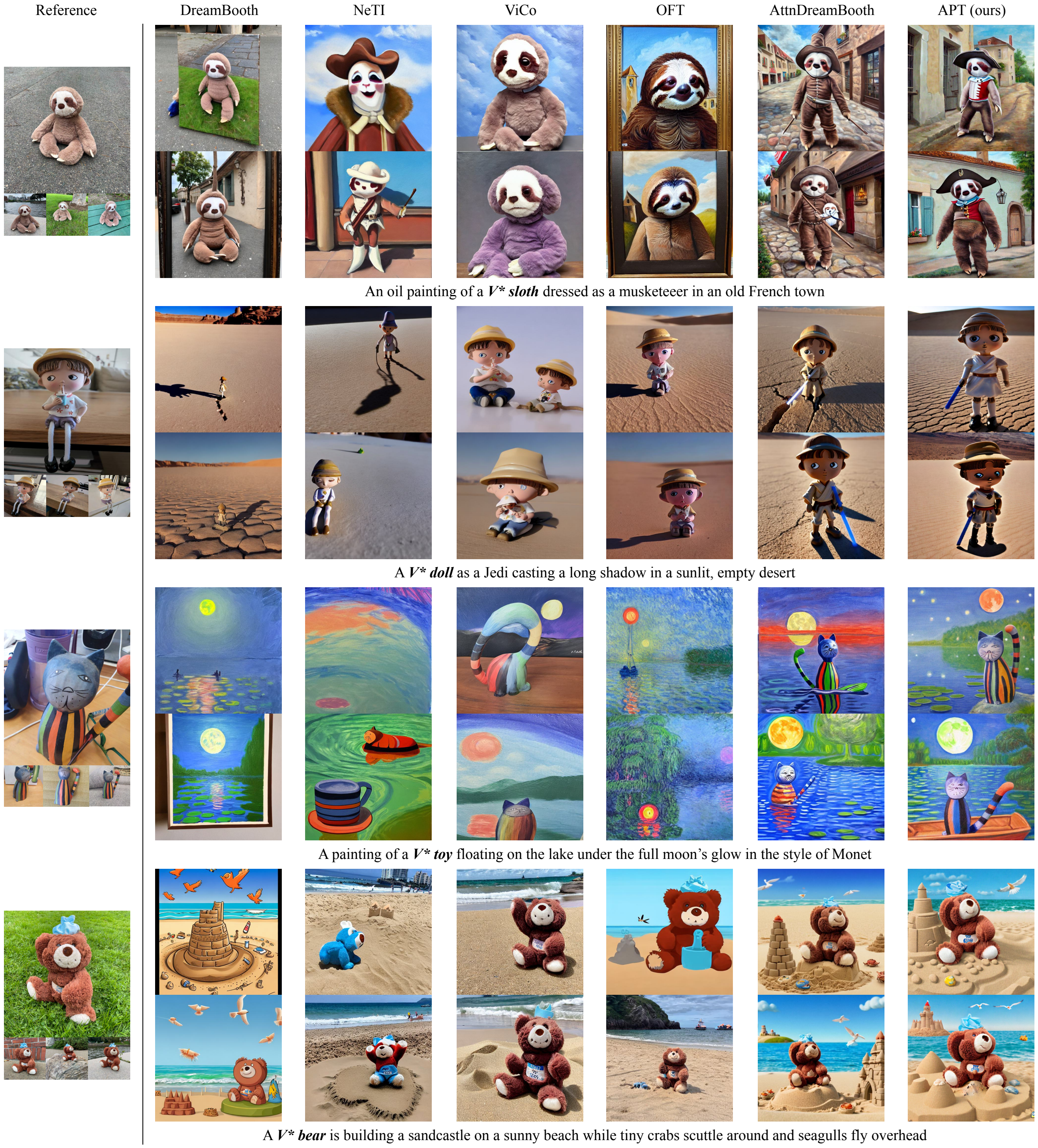
A ***sneaker*** sprinting on a running track, painted in impressionist style

A robot wearing ***sneaker***, wandering in the scrap heap

A ***backpack*** walking in a bustling marketplace surrounded by vibrant fruits and spices, in pop art style

A ***backpack*** collecting nuts in an autumn forest, illusrated in art nouveau style

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.



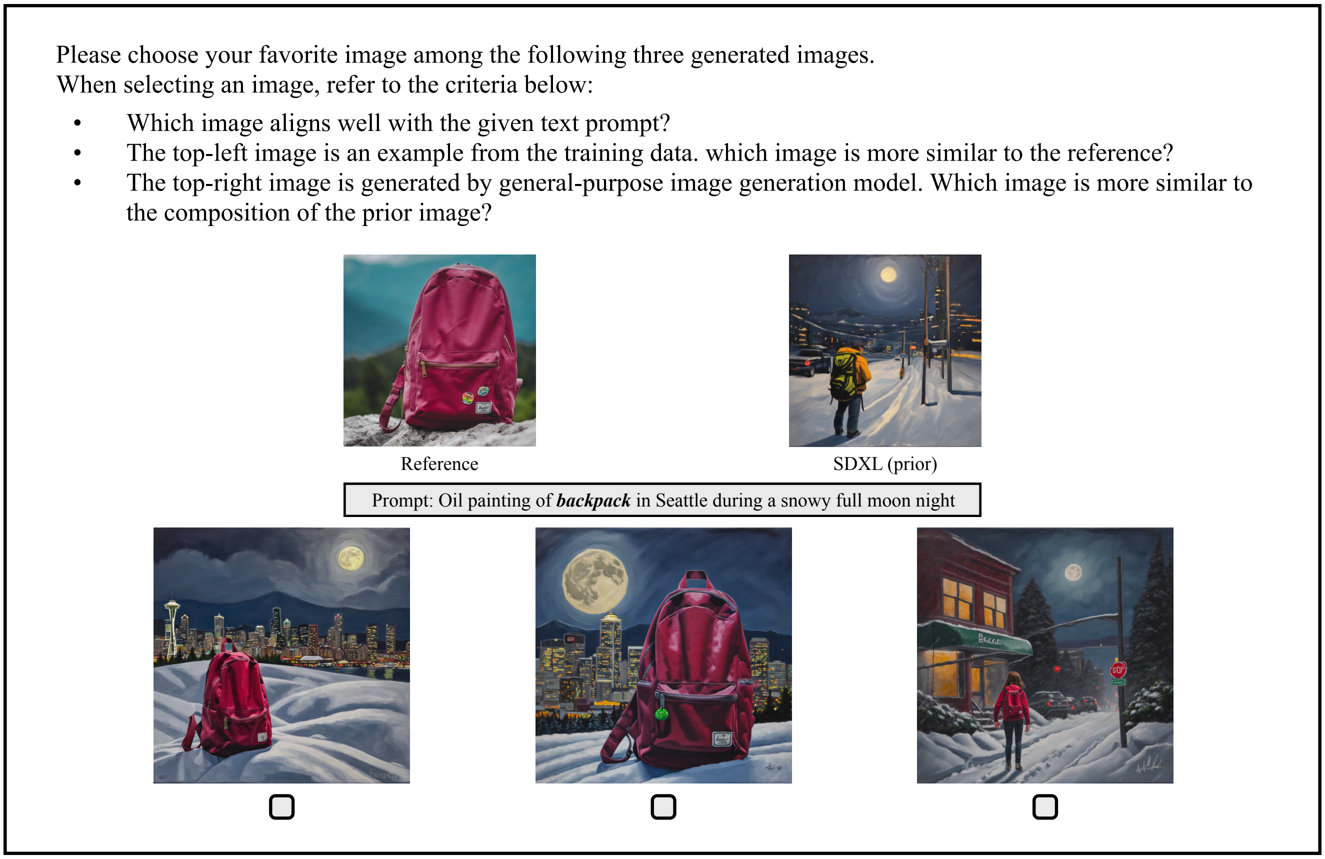
An oil painting of a ***V* sloth*** dressed as a musketeeer in an old French town

A ***V* doll*** as a Jedi casting a long shadow in a sunlit, empty desert

A painting of a ***V* toy*** floating on the lake under the full moon's glow in the style of Monet

A ***V* bear*** is building a sandcastle on a sunny beach while tiny crabs scuttle around and seagulls fly overhead

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.