# Powerful and Flexible: Personalized Text-to-Image Generation via Reinforcement Learning

Fanyue Wei[1], Wei Zeng[2], Zhenyang Li[2], Dawei Yin[2], Lixin Duan[1], and Wen Li[1†]

[1]University of Electronic Science and Technology of China
{wfanyue, wzeng316, zhenyounglee, lxduan, liwenbnu}@gmail.com
[2]Baidu Inc
{yindawei}@acm.org

**Abstract.** Personalized text-to-image models allow users to generate varied styles of images (specified with a sentence) for an object (specified with a set of reference images). While remarkable results have been achieved using diffusion-based generation models, the visual structure and details of the object are often unexpectedly changed during the diffusion process. One major reason is that these diffusion-based approaches typically adopt a simple reconstruction objective during training, which can hardly enforce appropriate structural consistency between the generated and the reference images. To this end, in this paper, we design a novel reinforcement learning framework by utilizing the deterministic policy gradient method for personalized text-to-image generation, with which various objectives, differential or even non-differential, can be easily incorporated to supervise the diffusion models to improve the quality of the generated images. Experimental results on personalized text-to-image generation benchmark datasets demonstrate that our proposed approach outperforms existing state-of-the-art methods by a large margin on visual fidelity while maintaining text-alignment. Our code is available at: https://github.com/wfanyue/DPG-T2I-Personalization.

**Keywords:** Personalized Text-to-Image Generation · Reinforcement Learning · Visual Fidelity

## 1 Introduction

Recent advances in text-to-image generation [30, 33, 36] exhibit the impressive ability to synthesize high-quality impressive images. Such models are robust and can generate images of diverse concepts in a wide variety of backgrounds and contexts. This opened a new area of research and innovation.

†Corresponding author.

However, these generation models are uncontrolled and lack the ability to synthesize customized concepts from personal lives. For instance, it is not possible to query with a prompt/image from your own pets, friends, or personal objects and modify their poses, locations, styles, or backgrounds.

To achieve such customization, existing approaches [11, 18, 34] utilize a controlled fine-tuning mechanism which allows the possibility of embedding new concepts into the pre-trained text-to-image diffusion model. For example, Text-Inversion [11] personalizes image generation by learning a unique textural identifier of the new concept from a given set of images during fine-tuning. Then, the fine-tuned model is able to generate new variations of the input concept using a prompt containing the learned identifier. Besides, DreamBooth [34] fine-tunes the entire diffusion model to learn the personalized concept instead. It is additionally regularized by the super-class images to preserve the class-specific priors. In addition, Custom Diffusion [18] proposes to fine-tune the key and value parameters in each cross-attention layer to enhance computational efficiency. However, all these diffusion-based methods are trained by a simple reconstruction objective step-by-step which can hardly enforce appropriate visual consistency between the generated image and the reference images.

To this end, we design a novel framework for the task of text-to-image(T2I) personalization via reinforcement learning which is facilitated with various objectives, differential or even non-differential. There are existing text-to-image generation methods using reinforcement learning by utilizing human feedback [7, 8, 10, 17, 44]. They usually use the policy gradient approach to incorporate aesthetic assessment or human preference as the reward for the general text-to-image generation to improve the image quality or text-alignment. While under the personalized setting, usually given only $4 \sim 6$ images depicting personalized concepts, it is hard to train an appropriate customized reward model. In our work, different from the existing reinforcement learning methods using human feedback reward for text-to-image generation, we explore several ways for text-to-image personalization to provide the suitable reward model for capturing the long-term visual consistency of the personalized subjects in diffusion model and rich supervision signals.

In this study, we introduce a versatile framework designed to support various forms of supervision for personalized text-to-image generation. Illustrated in Fig. 1, our framework utilizes the deterministic policy gradient (DPG) algorithm to fine-tune the diffusion models. This involves the incorporation of a specific differentiable reward function that considers personalized concepts. Based on this new framework, we further introduce two new losses to capture the long-term visual consistency for the text-to-image personalization task and enrich the supervision to improve the visual fidelity for personalization

Experimental results show that our proposed approach surpasses existing state-of-the-art methods on multiple personalized text-to-image generation benchmarks by a large margin to preserve visual fidelity.

In summary, Our main contributions are as follows:

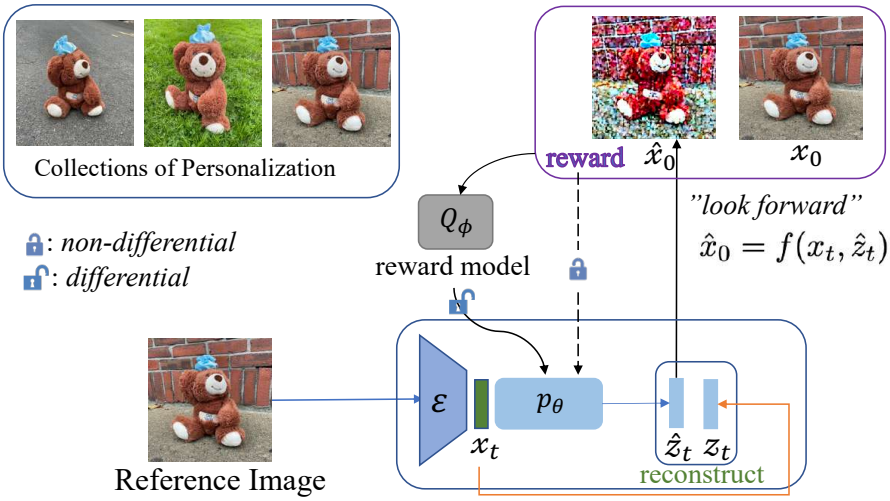



**Fig. 1:** Our proposed framework utilizes the DPG algorithm to capture the visual consistency and supervises the generation model with flexible objectives, differential or even non-differential.

- We design a novel framework for the task of text-to-image personalization via reinforcement learning. Especially, we regard the diffusion model as a deterministic policy and can be supervised by a learnable reward model for personalization.
- With the flexibility of our proposed framework, We introduce two new losses to improve the quality of generated images to capture the long-term visual consistency for the personalized details and enrich the supervision for the diffusion model.
- Experimental results on personalized text-to-image generation benchmarks demonstrate that our proposed approach surpasses existing state-of-the-art methods in visual fidelity.

## 2 Related Work

### 2.1 Diffusion Models for Image Generation

Diffusion-based image generation models [4, 9, 16, 26, 28, 30, 30, 31, 33, 36, 46] have developed rapid and impressive progress recently. DDPM [14] first performs a noise diffusion during the forward process and denoises on a Markov process. Then, DDIM [39] adopts an implicit estimation to accelerate the sample for image generation. As for text-to-image generation, there is also a huge progress. Imagen [36], GLIDE [27], Parti [45], Stable Diffusion [33] and DALL·E [4] have all exhibited impressive results on image generation given a textual prompt. In particular, Stable Diffusion [33] performs the diffusion process in the latent space, improving training and sampling efficiency significantly.

### 2.2 Personalized Text-to-Image Generation

Personalized text-to-image generation [2, 3, 11–13, 20, 34, 35] aims to adapt the pre-trained text-to-image generation model to learn a personalized concept from a given small set of images (*i.e.* $4 \sim 6$ images) and modify its pose, style or context.

Text Inversion [11] personalizes the image generation by learning a unique textural identifier of the new concept in given images during fine-tuning. Then, the fine-tuned model is able to generate new variations of the input concept using a prompt containing the learned identifier. $\mathcal{P}+$ [41] further improves the inversion method by injecting the learnable identifier into each attention layer of the denoising U-Net. In addition, NeTI [1] proposes to fuse the denoising process timestep on $\mathcal{P}+$ [41] by introducing a neural mapper.

By contrast, DreamBooth [34] fine-tunes the entire diffusion model to learn the personalized concept. It is regularized by the super-class images to preserve the class-specific priors. Custom-Diffusion [18] proposes to only fine-tune the key and value parameters in the cross-attention layers to enhance computational efficiency. ELITE [42] introduces to directly map the visual concepts into textual embeddings, by training a learnable encoder.

Besides, some works aim to provide a domain-specific text-to-image generator by utilizing a personalization encoder [2, 6, 12, 15, 22, 25, 37, 40]. Given a single image and a prompt, these models enable to generate images within a specific class domain without fine-tuning on new input images.

Differently, this paper revisits the task of personalized text-to-image generation via reinforcement learning and reforms the learning paradigm into a deterministic policy gradient (DPG) framework.

### 2.3 Reinforcement Learning for Text-to-image Generation

Reinforcement learning method [7, 17, 19, 21, 23, 32, 43, 44, 47] has been explored for text-to-image generation using human preference as reward, they usually use aesthetic assessment or human preference for a given prompt.

DPOK [10] finetunes text-to-image generation model using policy gradient with KL regularization using human feedback as reward, while DRaFT [8] relies on the differentiable reward to propagate the reward function gradient across the sampling procedure in the denoising process.

They all finetune the diffusion model for the general text-to-image generation. While for the personalized setting, usually given the only $4 \sim 6$ images, it is hard to train an appropriate reward model. In our work, we explore several ways for the personalized text-to-image generation task with the specific reward model for the given personalized concepts based on DPG [24, 38].

# 3 Method

## 3.1 Preliminaries

Stable Diffusion (SD) [33] is a latent text-to-image generation model based on DDPM [14]. It contains a large autoencoder $\mathcal{E}$ which is pretrained to extract latents from images, and a corresponding decoder $\mathcal{D}$ to map the latents back to images for reconstruction $\mathcal{D}(\mathcal{E}(I)) \approx I$. SD performs the diffusion process on the latent space of the autoencoder $(\mathcal{E}(\cdot), \mathcal{D}(\cdot))$. Then, the text conditions $y$ can be injected into the diffusion process by cross-attention. Thus, the training objective of the diffusion model is:

$$\mathcal{L}_{LDM} := \mathbb{E}_{x\sim\sigma(x),y,\epsilon\sim\mathcal{N}}\left[||\epsilon - \epsilon_\theta(x_t, t, \tau(y)||^2\right], \tag{1}$$

where $\mathcal{L}_{LDM}$ is a squared error loss, $\epsilon$ is the target noise, $\epsilon_\theta(\cdot)$ is a denoising network (*i.e.*, U-Net) to predict noise adding to the latents, $t$ is timestep in diffusion process, $x_t$ is the noisy latents in timestep $t$, and $\tau(\cdot)$ is the pretrained CLIP [29] text encoder in Stable Diffusion [33]. During inference, a random Gaussian noise $x_T \sim \mathcal{N}(0, 1)$ is iterative denoised to $x_0$, and the final image is obtained through the decoder $\hat{I}_0 = \mathcal{D}(x_0)$.

## 3.2 DPG framework for T2I Personalization

Existing approaches [11, 18, 34] for text-to-image personalization follows the training procedure presented in Sec. 3.1. The reconstruction loss is calculated during the diffusion process for $x_t$, thus cannot be directly used to optimize the final generation results $x_0$ by the visual details. Whereas RL technique can act as a flexible tool for optimization, and has achieved huge success in various fields due to its flexibility and powerful modeling capabilities. Inspired by this, we revisit the task of text-to-image personalization via reinforcement learning method. In particular, we treat the diffusion model as a deterministic policy and propose a flexible framework that can facilitate various supervision for personalized text-to-image generation with a learnable specific reward.

Next, we outline the main formulation of our DPG framework. The deterministic policy applies the action that maximizes the Q-function $Q_\phi(\cdot)$, and the Q-function is assumed to be differentiable with respect to the action. In our framework, We regard the latent state, the timestep, and the encoded text condition $\{x_t, t, \tau(y)\}$ as the input, the predicted noise $z_t$ as the action, and the text-to-image generation model denoted by $\epsilon_\theta(x_t, t, \tau(y))$ as the policy. We define the policy function as follows,

$$\hat{z}_t = \epsilon_\theta(x_t, t, \tau(y)). \tag{2}$$

At each timestep, the policy model $\epsilon_\theta(x_t, t, \tau(y))$ takes the latent $x_t$ of current timestep $t$ and the text condition $y$ as input and generates the action $\hat{z}_t$ during the training process.

As shown in Eq. 3, the optimization purpose of the DPG framework is to maximize the expectation of the accumulated reward.

$$\max_{\theta} \mathbb{E}\left[Q_\phi(x_t, \epsilon_\theta(x_t, t, \tau(y)))\right], \tag{3}$$

where $Q_\phi(x_t, \epsilon_\theta(x_t, t, \tau(y)))$ aims to calculate the cumulative reward when applies the action $z_t$ at the state $\{x_t, \tau(y), t\}$, and its implementation is very flexible.

We optimize the Q-network to predict the accumulative reward. To achieve this objective, $Q_\phi(\cdot)$ is updated by the gradient descent algorithm using Eq. 4 concurrently.

$$\min_{\phi} ||Q_\phi(x_t, \epsilon_\theta(x_t, t, \tau(y))) - \sum_{i=0}^{t} r(x_i, \tau(y), i)||^2, \tag{4}$$

where $r(x_i, \tau(y), i)$ denotes the reward on the timestep $i$.

While in the diffusion and denoising process, the policy model $\epsilon_\theta(\cdot)$ is optimized to minimize Eq. 1 in one timestep. Therefore, minimizing the reconstruction loss encourages policy model $\epsilon_\theta(\cdot)$ to make rewards (*i.e.*, minimizing Eq. 1 equals to maximizing Eq. 3 for diffusion model). In this case, the reward function is obtained as follows,

$$r(x_t, t, \tau(y)) = -||\epsilon - \epsilon_\theta(x_t, t, \tau(y))||^2. \tag{5}$$

$Q_\phi(\cdot)$ works to directly predict the one step immediate reward in Eq. 5 for the diffusion model $\epsilon_\theta(\cdot)$ given $x_t$ to estimate the noise in timestep $t-1$.

Therefore, as shown in Fig. 1, the entire DPG framework for text-to-image personalization can be optimized by Eq. 4 to train the diffusion model $\epsilon_\theta(\cdot)$ concurrently with $Q_\phi(\cdot)$. The algorithm pseudo code is presented in Algorithm 1.

### 3.3 Learning to "Look Forward"

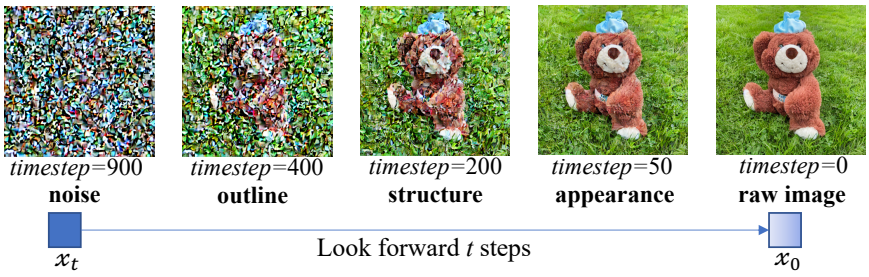



**Fig. 2:** During the denoising process, in the early timesteps ($t \approx T$), the diffusion model attempts to represent the outline and structure of the subject, whereas in the later steps ($t \approx 0$), the model focuses on the visual details.

**Algorithm 1** DPG Framework for T2I Personalization

Input: Policy function $\epsilon_\theta(\cdot)$, Q-function $Q_\phi(\cdot)$, the timestep $T$ and the text condition $y$
**repeat**
  Randomly choose an image from the set of reference images
  Randomly sample $t \sim \{0, \cdots, T-1\}$
  Process $t$ step diffusion process, obtain the latent state $x_t$
  Calculate the reward $r$ based on the definition
  Obtain the accumulative reward as the target

$$\sum_{i=0}^{t} r(x_i, \tau(y), i)$$

  Update the Q-function parameters $\phi$ by one step of gradient descent using

$$\nabla_\phi ||Q_\phi(x_t, \epsilon_\theta(x_t, t, \tau(y))) - \sum_{i=0}^{t} r(x_i, \tau(y), i)||^2$$

  Update the diffusion model $\theta$ by gradient ascent using

$$\nabla_\theta Q_\phi(x_t, \epsilon_\theta(x_t, t, \tau(y)))$$

**until** End Training

Existing methods [11, 18, 34] follow the paradigm that the diffusion model is optimized by the reconstruction loss during the diffusion process step-by-step, thus cannot be directly optimized with the final generation images by the visual details. However, as shown in Fig. 2 intuitively, the denoising process is guided by different nature implicitly. At the different timesteps, the generation model focuses on the different features, in the early timesteps ($t \approx T$), the diffusion model attempts to recognize the outline of the subject and determine the structure of the subject, while in the later steps ($t \approx 0$), the model focuses on the visual fine details. Thus, we argue for taking advantage of the "look forward" of reinforcement learning to implicitly guide the generation model to capture the long-term visual consistency. This approach encourages the generation model to focus on the different features at different timesteps, improving the visual consistency of the personalized subject.

Since the reward in Eq. 5 represents the one-step direct reward, we aim to leverage the nature of the diffusion process to "look forward" to $\hat{x}_{0,t}$. In the diffusion process, the Gaussian noise $z_t$ is added into the initial latent $x_0$ at timestep $t$ as follows,

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} z_t. \tag{6}$$

During the denoising process, the policy diffusion model $\epsilon_\theta(z_t, t, \tau(y))$ aims to estimate the noise $\hat{z}_t$ from $x_t$. Consequently, we can obtain the $\hat{x}_{0,t}$ in Eq. 7, derived from Eq. 6 at the given timestep $t$.

$$\hat{x}_{0,t} = \frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1-\bar{\alpha}_t}\hat{z}_t). \tag{7}$$

After obtaining the final generation results $\hat{x}_{0,t}$, our purpose can transition from the step-by-step reconstruction of the diffusion noise to direct comparison of final generation results between $x_0$ and $\hat{x}_{0,t}$ at timestep $t$.

To achieve the objective of "looking forward" to implicitly guide the focus at different denoising states, the reward function in Eq. 5 can be rewritten between $x_0$ and $\hat{x}_{0,t}$ as follows,

$$\begin{aligned}\mathcal{L} &= -||\hat{x}_{0,t} - x_0||^2 \\ &= -||\frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1-\bar{\alpha}_t}\hat{z}_t) - \frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1-\bar{\alpha}_t}z_t)||^2 \\ &= -\frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}||\hat{z}_t - z_t||^2.\end{aligned} \tag{8}$$

With the optimization objective on $\hat{x}_{0,t}$ in Eq. 8, our DPG framework can learn to look forward from $x_t$ to $\hat{x}_{0,t}$ to acquire the implicit guidance at different timestep $t$ to enforce appropriate long-term structural consistency between the generated image and the reference images. Thus, the reward function in Eq. 8 can be used to update $Q_\phi(\cdot)$ at step 8 in Algorithm 1.

In addition, to make utilization of the prior that the Gaussian noise is added gradually over $t$ steps and denoised step-by-step, we accumulated the reward from $t$ to 0 in Eq. 9 to make consistency with the denoising process rather than calculating one-step direct rewards,

$$Q_\phi(x_t, \epsilon_\theta(x_t, t, \tau(y))) = \mathcal{L}(x_t, \tau(y), t) + \gamma Q_\phi(x_{t-1}, \epsilon_\theta(x_{t-1}, t-1, \tau(y))). \tag{9}$$

The pseudo code is included in the supplementary materials.

### 3.4 Learning Complex Reward

With our proposed DPG framework equipped with "Look Forward" for text-to-image personalization, we are capable to incorporate various complex objectives, differential or even non-differential, to learn a specific reward model $Q_\phi(\cdot)$ to supervise the generation models to improve the quality of generated images. As shown in Fig. 3, the aforementioned strengths arise from two aspects: one is that "Look forward" allows us to optimize the model based on the final generation results, other is that the learnable $Q_\phi(\cdot)$ for personalized subjects facilities complex supervision.

As the self-supervised learning method DINO [5] encourages the unique visual features for personalization [34], in this work, we select the DINO similarity as the representative reward function.

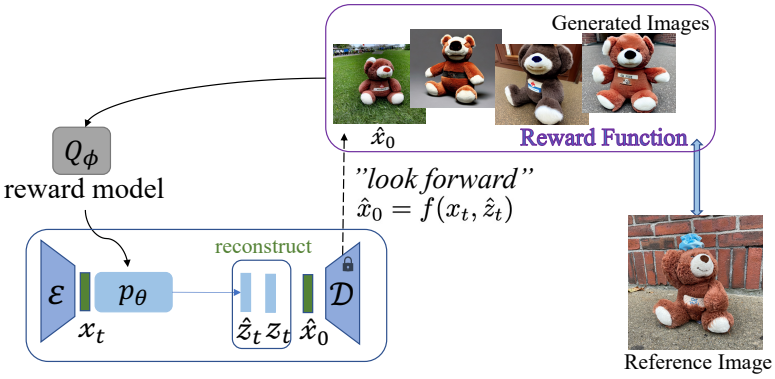



**Fig. 3:** Our proposed framework of DPG equipped with "looking forward" can further introduce more flexible supervision with a learnable reward model for the personalized generation model (*e.g.*, Stable Diffusion).

With our specific reward model $Q_\phi(\cdot)$ for the given collection of personalized reference images $I(\cdots)$, we can simply incorporate the DINO reward for the diffusion model in our DPG framework.

To adopt the DINO similarity as reward, we utilize $\mathcal{D}(\cdot)$ to decode the final generation results $\hat{x}_{0,t}$ to the final generated image $\hat{I}$ by using $\hat{I} = \mathcal{D}(\hat{x_{0,t}})$. Then, we obtain the image embeddings from DINO image encoder $\mathcal{X}(\cdot)$ by $\kappa = \mathcal{X}(I)$. Thus, the reward function $r(\cdot)$ to be estimated by $Q_\phi(\cdot)$ can be formulated as follows,

$$r(x_t) = -(1 - \hat{\kappa} \cdot \kappa), \tag{10}$$

where $\hat{\kappa}$ is embeddings of the generated image extracted from $\mathcal{X}(\cdot)$, while $\kappa$ refers to the embeddings of the reference image.

Then, we can inject the supervision of unique visual features of the personalized subjects with the original reconstruction reward, as in Eq. 5. Thus we adopt the combination of both the DINO reward and reconstruction reward in Eq. 5 to work as the complex reward function, controlled with a weight $\lambda$, where $B$ denotes batch size. The gradient at step 9 in Algorithm 1 can be refined as follows,

$$\nabla_\theta \frac{1}{B} \sum_B (\lambda Q_\phi(x_t, \epsilon_\theta(x_t, t, \tau(y))) + (-||\epsilon - \epsilon_\theta(x_t, t, \tau(y))||^2)). \tag{11}$$

With our flexible DPG framework for reward, the reward can be differential to any other metric that can reflect on the generated image along with the reference personalized images.

# 4 Experiments

## 4.1 Experimental Setup

**Datasets.** We adopt the DreamBooth benchmark [34] to evaluate our DPG framework for text-to-image personalization. The dataset comprises 30 concepts across 15 different categories. Among these, 9 subjects belong to live pets (*i.e.*, dogs and cats). The remaining 21 subjects pertain to various objects such as backpacks, cars and *etc*. The dataset contains $4 \sim 6$ of images per concept, each captured under differing conditions, in various environments, and from multiple perspectives. Moreover, the dataset also contains 25 challenging prompts to evaluate the text-to-image personalization methods. Besides, we also conduct experiments on Custom benchmark [18].

**Table 1:** Quantitative comparisons with existing methods

| Method | DINO | CLIP-I | CLIP-T |
|---|---|---|---|
| Custom Diffusion [18] | 0.649 | 0.712 | 0.321 |
| Custom Diffusion w/ Our DINO reward | 0.640 | 0.715 | 0.320 |
| Custom Diffusion w/ Our Look Forward | 0.669 | 0.728 | **0.322** |
| DreamBooth [34] | 0.694 | 0.762 | 0.282 |
| DreamBooth w/ Our DINO reward | 0.723 | 0.783 | 0.270 |
| DreamBooth w/ Our Look Forward | **0.738** | **0.797** | 0.269 |

**Evaluation:** Following existing text-to-image personalization methods [11, 18, 34], we evaluate our proposed approach using *Image-Alignment* and the *Text-Alignment*. The *Image-Alignment* measures the subject fidelity in the generated images while the *Text-Alignment* aims to evaluate the similarity between the generated images and the given prompt.

For *Image-Alignment*, we adopt DINO [5] and CLIP-I [29]. The objective of self-supervised training in DINO [5] is to encourage the discrimination of unique features of the subject, while CLIP-I [29] may focus on the semantic feature space such as color. Both DINO [5] and CLIP-I [29] calculate the average pairwise cosine similarity between the extracted embeddings by the image encoder of the generated image and the reference image. For the *Text-Alignment*, we calculate the similarity of the CLIP embeddings between the textual prompt and the generated image. Following existing methods [5, 18, 34], we adopt ViT-B/32 for the CLIP model and ViT-S/16 for the DINO model to extract visual and textual features.

**Implementation Details**. We adopt DreamBooth [34] as our baseline method. Since the official code of DreamBooth [34] is not publicly available, we use the implementation in the popular "diffusers" library and the pretrained Stable Diffusion V1.4 for all compared methods for fair comparison. For Custom Diffusion [18], we reproduce their methods using their official code and the default

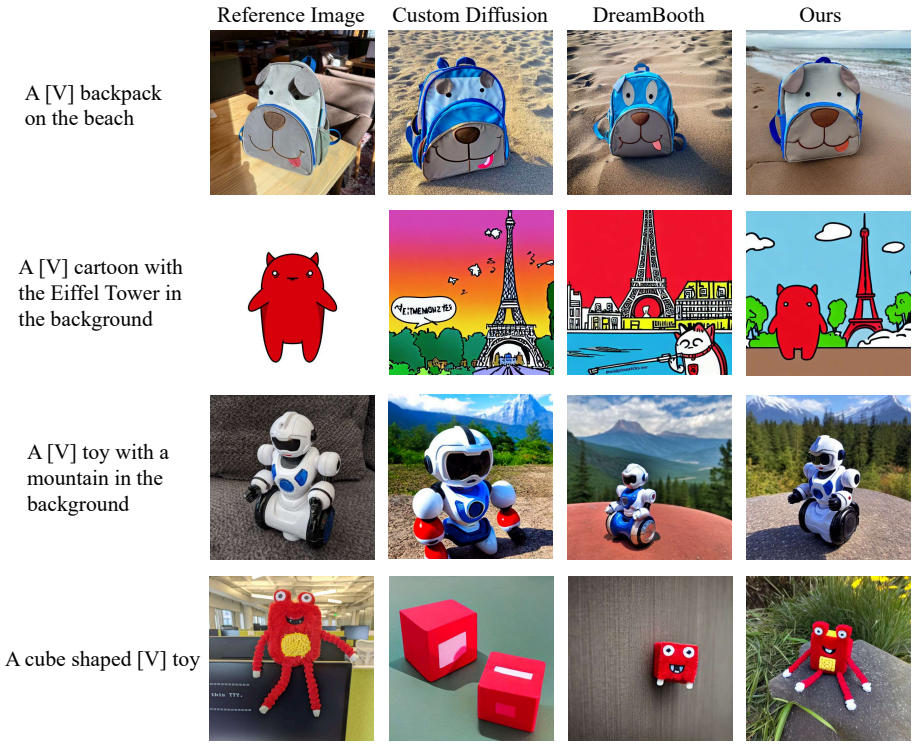



**Fig. 4:** In this figure, we present the reference images alongside the images generated by Custom Diffusion, DreamBooth and our method. As demonstrated, given the challenging textual prompts, the images generated by Ours best preserve the high fidelity of the personalized attributes, including color, expressions, texture and *etc.*

hyperparameters on the DreamBooth benchmark. The resolution of generated images is $512 \times 512$. All experiments are performed on a 32G V100 card. More implementation details are included in the supplementary materials.

### 4.2 Qualitative Results.

We present the visualization results for qualitative comparisons. As shown in Fig. 4, our method is capable to capture the visual details (*i.e.*, color, texture, poses and *etc.*) of the personalized subjects with various prompts while following the textual prompts faithfully. For example, the generated images by *Ours* best capture the color of the reference images for the "backpack" and the shape of "cartoon" than the compared methods. The qualitative results demonstrate that our methods achieve better fidelity while preserving text-alignment. Additionally, our flexible DPG framework allows our method to be easily extended to other rewards for corresponding purposes. More visualization examples are included in the supplementary materials.

### 4.3 Quantitative Results

To validate the effectiveness of our proposed DPG framework, we compare our approaches (including the "Look Forward" reconstruction reward and the DINO reward) with several state-of-the-art methods. The compared methods include Custom Diffusion [18] and DreamBooth [34]. As illustrated in Table 1 and 2, our methods achieve the highest image-alignment on both DINO and CLIP-I evaluation metrics while preserving the text-alignment on CLIP-T metric.

Especially, DreamBooth equipped with our proposed DPG framework improves the visual fidelity by a large margin 2.1% on DINO and 1.8% on CLIP-I than the DreamBooth baseline methods. For text-alignment, there is an intrinsic trade-off between the text-alignment and image-alignment [18]. The current challenge of text-to-image personalization lies in improving visual consistency with the reference images, while the maintenance of text-alignment is primarily handled by the base T2I generation model such as Stable Diffusion and the text encoder. Our proposed methods outperform the compared methods on visual fidelity and preserve the text-alignment. For the Custom Diffusion [18] baseline, this approach only fine-tunes the parameters of cross-attention layers (between the text and image) while not emphasizing the visual embeddings, thus achieving better text-alignment performance yet poor image-alignment than the compared methods of the DreamBooth baseline.

**User Study.** We conduct a user study to compare our proposed approach with DreamBooth [34] to evaluate the human preference with the generation results for both the image-alignment and text-alignment. We provide three randomly selected personalized datasets for the user study. The participants are required to compare the generated images, which are the best results selected from eight images generated by different methods with the same random seed, given both the prompt and reference images. For the image fidelity, the participants are required to choose which method best preserves personalized visual consistency with reference images and which is most consistent with the prompt for text-alignment. As illustrated by Table 3, our method preserves image fidelity better than the compared method by a large margin while achieving comparable performance of text fidelity, which indicates that our approach can generate images that are more appealing to human preference.

**Computation Cost.** Our lightweight reward model operates on the latent space, thus introducing negligible computational cost. The number of trainable parameters is $0.26M$ *versus* $859.40M$ trainable U-Net parameters.

Besides, we provide more analysis in the supplementary materials.

### 4.4 Ablation Studies

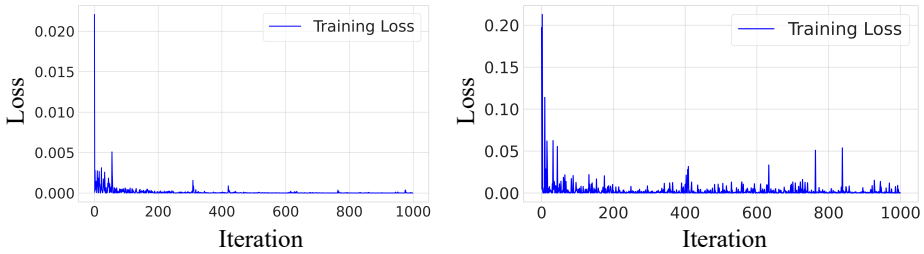
We conducted ablation studies to verify the different components of our DPG framework, including the discount rate and the weight of the dino-reward. In addition, Fig. 5 illustrates that our proposed reward model is easy to converge for both "Look Forward" reward and DINO reward.

**Table 2:** Evaluation on Custom Benchmark.

| Method | DINO | CLIP-I | CLIP-T |
|---|---|---|---|
| DreamBooth | 0.640 | 0.737 | 0.309 |
| DreamBooth w/ LF | **0.680** | **0.773** | 0.303 |
| DreamBooth w/ DINO | 0.653 | 0.753 | **0.310** |

**Table 3:** Quantitative comparisons of User Preference

| Method | Ours | DreamBooth | Similar |
|---|---|---|---|
| Image Fidelity | 55.1% | 12.0% | 32.9% |
| Text Fidelity | 19.6% | 20.4% | 60.0% |




(a) Convergence of Look Forward Reward (b) Convergence of DINO reward

**Fig. 5:** The convergence of the Q-function is illustrated in the subfigures. Subfigure (a) presents the training loss of the Q-function for the reconstruction reward, while subfigure (b) relates to the DINO reward.

**Effectiveness of the discount rate:** We adapt the discount rate of reinforcement learning to verify the sensitivity of our framework. We use different $\gamma$ on the "clock" collection to evaluate the robustness of our method. As shown in Tab. 4, even with different discount rate, our methods still maintain high fidelity.

**Table 4:** Sensitivity of discount rate $\gamma$

| $\gamma$ | DINO | CLIP-I | CLIP-T |
|---|---|---|---|
| DB [34] | 0.644 | 0.707 | 0.239 |
| w/o $\gamma$ | 0.727 | 0.761 | 0.209 |
| $\gamma = 0.9986$ | 0.704 | 0.743 | 0.213 |

**Table 5:** Sensitivity of weight $\lambda$

| $\lambda$ | DINO | CLIP-I | CLIP-T |
|---|---|---|---|
| DB [34] | 0.644 | 0.707 | 0.239 |
| $\lambda = 0.1$ | 0.704 | 0.743 | 0.213 |
| $\lambda = 1$ | 0.727 | 0.746 | 0.211 |

**Different weight for DINO reward $\lambda$:** We evaluate the sensitivity of our DPG framework by ablating the weight $\lambda$ for the DINO reward. We conduct experiments using different $\lambda$ values of 0.1 and 1 on the "robot_toy" dataset to demonstrate the robustness of our approach. As shown in Table 5, increasing the weight $\lambda$ for the DINO reward from 0.1 to 1 improves the visual fidelity marginally. The DINO metric steeply increases from 0.644 to 0.727, and CLIP-I from 0.707 to 0.746. However, a trade-off may exist between the high visual reconstruction of personalization and text-alignment. As shown in Table 5, CLIP-T drops from 0.239 to 0.211. This indicates that as the DINO reward increases,


**Fig. 6:** As illustrated in the figure, with the textual prompt "A [V] toy floating on the water" to generate images, increasing the weight of DINO reward may preserve the attributes of the personalized subject but damage the ability of text-alignment.

the generation model tends to emphasize visual fidelity, which may potentially compromise the text-alignment ability. We present the images generated by the generation model with two different $\lambda$ weights in Fig. 6, the robot generated with higher $\lambda$ overemphasizes better visual consistency with the reference images but follows the textual prompt less faithfully.

## 5 Conclusion

We design a novel framework for text-to-image personalization via reinforcement learning. Especially, we treat the diffusion model as a deterministic policy that can be supervised by a learnable reward model for personalization. With the flexibility of our framework, we introduce two new losses to improve the quality of generated images. The proposed method is capable to capture the long-term visual consistency of personalized details and enrich the supervision of the diffusion model. Experiments on several benchmarks demonstrate that our approach surpasses existing methods in visual fidelity while preserving text-alignment.
**Limitations:** In some cases, our framework equipped with such baselines (*e.g.*, DreamBooth) may overemphasize the visual fidelity. The issue can be alleviated with a stronger text encoder or by resorting to baselines which balance the alignment between image and text. Moreover, we will further design the text-alignment related reward with our DPG framework to improve text-alignment.
**Social Impact:** Our methods can synthesize some fake images with personalized subjects such as human face or private pets, which may increase the risk of privacy leakage and portrait forgery. Therefore, users intending to use our technique should apply for authorization to use the respective personalized images. Nevertheless, our approach can serve as a tool for AIGC to create imaginative images for entertainment purposes.

## Acknowledgement

This work is supported by the National Natural Science Foundation of China (No. 62176047), the Sichuan Science and Technology Program (No. 2022YFS0600), and the Fundamental Research Funds for the Central Universities Grant (No. ZYGX2021YGLH208).

## References


1. Alaluf, Y., Richardson, E., Metzer, G., Cohen-Or, D.: A neural space-time representation for text-to-image personalization. ACM Transactions on Graphics (TOG) **42**(6), 1–10 (2023)
2. Arar, M., Gal, R., Atzmon, Y., Chechik, G., Cohen-Or, D., Shamir, A., H. Bermano, A.: Domain-agnostic tuning-encoder for fast personalization of text-to-image models. In: SIGGRAPH Asia 2023 Conference Papers. pp. 1–10 (2023)
3. Arar, M., Voynov, A., Hertz, A., Avrahami, O., Fruchter, S., Pritch, Y., Cohen-Or, D., Shamir, A.: Palp: Prompt aligned personalization of text-to-image models. arXiv preprint arXiv:2401.06105 (2024)
4. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., Manassra, W., Dhariwal, P., Chu, C., Jiao, Y.: Dall-e 3. `https://cdn.openai.com/papers/dall-e-3.pdf` (2023)
5. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
6. Chen, W., Hu, H., Li, Y., Ruiz, N., Jia, X., Chang, M.W., Cohen, W.W.: Subject-driven text-to-image generation via apprenticeship learning. Advances in Neural Information Processing Systems **36** (2024)
7. Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S., Amodei, D.: Deep reinforcement learning from human preferences. Advances in neural information processing systems **30** (2017)
8. Clark, K., Vicol, P., Swersky, K., Fleet, D.J.: Directly fine-tuning diffusion models on differentiable rewards. In: The Twelfth International Conference on Learning Representations (2023)
9. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
10. Fan, Y., Watkins, O., Du, Y., Liu, H., Ryu, M., Boutilier, C., Abbeel, P., Ghavamzadeh, M., Lee, K., Lee, K.: Reinforcement learning for fine-tuning text-to-image diffusion models. Advances in Neural Information Processing Systems **36** (2024)
11. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. In: The Eleventh International Conference on Learning Representations (2022)
12. Gal, R., Arar, M., Atzmon, Y., Bermano, A.H., Chechik, G., Cohen-Or, D.: Encoder-based domain tuning for fast personalization of text-to-image models. ACM Transactions on Graphics (TOG) **42**(4), 1–13 (2023)
13. Hao, S., Han, K., Zhao, S., Wong, K.Y.K.: Vico: Detail-preserving visual condition for personalized text-to-image generation. arXiv preprint arXiv:2306.00971 (2023)
14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
15. Jia, X., Zhao, Y., Chan, K.C., Li, Y., Zhang, H., Gong, B., Hou, T., Wang, H., Su, Y.C.: Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. arXiv preprint arXiv:2304.02642 (2023)
16. Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6007–6017 (2023)

17. Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O.: Pick-a-pic: An open dataset of user preferences for text-to-image generation. Advances in Neural Information Processing Systems **36** (2024)
18. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1931–1941 (2023)
19. Lee, K., Liu, H., Ryu, M., Watkins, O., Du, Y., Boutilier, C., Abbeel, P., Ghavamzadeh, M., Gu, S.S.: Aligning text-to-image models using human feedback. arXiv preprint arXiv:2302.12192 (2023)
20. Lee, K., Kwak, S., Sohn, K., Shin, J.: Direct consistency optimization for compositional text-to-image personalization. arXiv preprint arXiv:2402.12004 (2024)
21. Lee, S.H., Li, Y., Ke, J., Yoo, I., Zhang, H., Yu, J., Wang, Q., Deng, F., Entis, G., He, J., et al.: Parrot: Pareto-optimal multi-reward reinforcement learning framework for text-to-image generation. arXiv preprint arXiv:2401.05675 (2024)
22. Li, D., Li, J., Hoi, S.: Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. Advances in Neural Information Processing Systems **36** (2024)
23. Liang, Y., He, J., Li, G., Li, P., Klimovskiy, A., Carolan, N., Sun, J., Pont-Tuset, J., Young, S., Yang, F., et al.: Rich human feedback for text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19401–19411 (2024)
24. Lillicrap, T.P., Hunt, J.J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., Wierstra, D.: Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971 (2015)
25. Ma, J., Liang, J., Chen, C., Lu, H.: Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. arXiv preprint arXiv:2307.11410 (2023)
26. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning. pp. 8162–8171. PMLR (2021)
27. Nichol, A.Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Mcgrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In: International Conference on Machine Learning. pp. 16784–16804. PMLR (2022)
28. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
29. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
30. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
31. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International Conference on Machine Learning. pp. 8821–8831. PMLR (2021)
32. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7008–7024 (2017)

33. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
34. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22500–22510 (2023)
35. Ruiz, N., Li, Y., Jampani, V., Wei, W., Hou, T., Pritch, Y., Wadhwa, N., Rubinstein, M., Aberman, K.: Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. arXiv preprint arXiv:2307.06949 (2023)
36. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022)
37. Shi, J., Xiong, W., Lin, Z., Jung, H.J.: Instantbooth: Personalized text-to-image generation without test-time finetuning. arXiv preprint arXiv:2304.03411 (2023)
38. Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., Riedmiller, M.: Deterministic policy gradient algorithms. In: International conference on machine learning. pp. 387–395. Pmlr (2014)
39. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
40. Tewel, Y., Gal, R., Chechik, G., Atzmon, Y.: Key-locked rank one editing for text-to-image personalization. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–11 (2023)
41. Voynov, A., Chu, Q., Cohen-Or, D., Aberman, K.: $p+$: Extended textual conditioning in text-to-image generation. arXiv preprint arXiv:2303.09522 (2023)
42. Wei, Y., Zhang, Y., Ji, Z., Bai, J., Zhang, L., Zuo, W.: Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. arXiv preprint arXiv:2302.13848 (2023)
43. Wu, X., Sun, K., Zhu, F., Zhao, R., Li, H.: Better aligning text-to-image models with human preference. arXiv preprint arXiv:2303.14420 (2023)
44. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: Imagereward: Learning and evaluating human preferences for text-to-image generation. arXiv preprint arXiv:2304.05977 (2023)
45. Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B.K., et al.: Scaling autoregressive models for content-rich text-to-image generation. arXiv preprint arXiv:2206.10789 **2**(3), 5 (2022)
46. Zhang, S., Xiao, S., Huang, W.: Forgedit: Text guided image editing via learning and forgetting. arXiv preprint arXiv:2309.10556 (2023)
47. Zhang, Y., Tzeng, E., Du, Y., Kislyuk, D.: Large-scale reinforcement learning for diffusion models. arXiv preprint arXiv:2401.12244 (2024)