InstructBooth: Instruction-following Personalized Text-to-Image Generation

Daewon Chae [1] **Nokyung Park** [2] **Jinkyu Kim** [* 1] **Kimin Lee** [* 3]

Abstract

Personalizing text-to-image models using a limited set of images for a specific object has been explored in subject-specific image generation. However, existing methods often face challenges in aligning with text prompts due to overfitting to the limited training images. In this work, we introduce InstructBooth, a novel method designed to enhance image-text alignment in personalized text-to-image models without sacrificing the personalization ability. Our approach first personalizes text-to-image models with a small number of subject-specific images using a unique identifier. After personalization, we fine-tune personalized text-to-image models using reinforcement learning to maximize a reward that quantifies image-text alignment. Additionally, we propose complementary techniques to increase the synergy between these two processes. Our method demonstrates superior image-text alignment compared to existing baselines, while maintaining high personalization ability. In human evaluations, InstructBooth outperforms them when considering all comprehensive factors.

1. Introduction

Recently, text-to-image models (Ramesh et al., 2022; Saharia et al., 2022; Rombach et al., 2022) have demonstrated superior performance in generating natural and high-quality images given novel text prompts. These models can produce photorealistic images of general objects in diverse contexts using a natural language prompt. Based on these advancements, the new research question arises: *How can we enable text-to-image models to generate personalized subject images?* For example, given only a few images of an Olympic mascot plushie, the goal is to make models generate images featuring this plushie in various contexts, such as participating in Olympic sports at the stadium (as
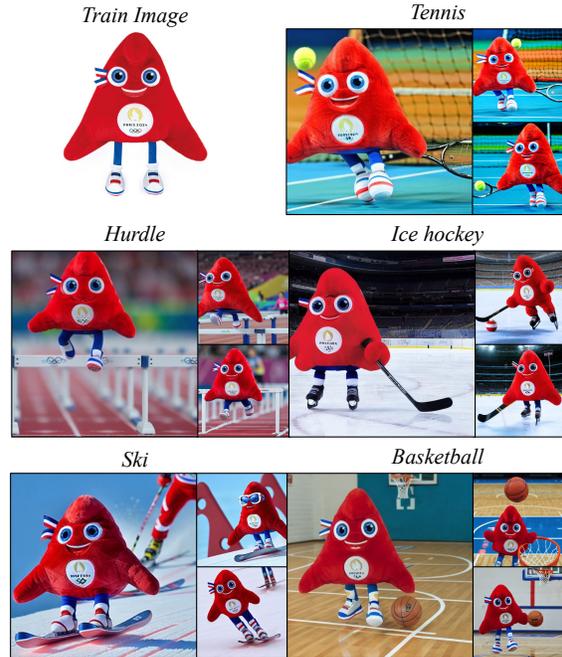

Figure 1. We propose InstructBooth, a method that enables the generation of images featuring specific user-provided subjects without degradation in image-text alignment. For example, InstructBooth can create new images of *unseen* Phryge, the Paris 2024 Olympic mascot plushie, participating in various sports.

illustrated in Figure 1). This capability holds the potential to open up exciting possibilities in personalized image generation, empowering users to effortlessly create custom imagery tailored to their specific interests and preferences.

To personalize existing text-to-image models, several approaches have been proposed that learn user-defined concepts using a few given images (Ruiz et al., 2023; Gal et al., 2022; Voynov et al., 2023; Alaluf et al., 2023; Zhang et al., 2023; Liu et al., 2023). Despite the promise of prior methods, they often exhibit issues with low text fidelity due to overfitting on the limited training images. As shown in Figure 3, when provided with the text prompt "playing soccer", DreamBooth (Ruiz et al., 2023) becomes overfitted and fails to reflect the desired action which input prompt requires. To mitigate this overfitting issue, several studies (Kumari et al., 2023; Han et al., 2023; Tewel et al., 2023) have proposed the personalization method that constrains the trainable weights of pre-trained text-to-image models. While this approach

*Equal advising [1]Korea University [2]NAVER [3]KAIST. Correspondence to: Jinkyu Kim <jinkyukim@korea.ac.kr>, Kimin Lee <kiminlee@kaist.ac.kr>.

shows some improvement in text fidelity, it often results in significant degradation of personalization capability. In the case of Custom Diffusion (Kumari et al., 2023), which only fine-tunes specific layers, the generated images are notably less similar to the reference images compared to those produced by DreamBooth, as demonstrated in Figure 3. Therefore, given the limitations of recent approaches, enhancing both subject and text fidelity remains a significant challenge in the personalization task.

In this work, we aim to achieve both high subject personalization and text fidelity by addressing the challenges associated with supervised learning using a limited dataset of images. Our main idea is to introduce reinforcement learning (RL) fine-tuning in a subsequent stage. Specifically, we first personalize text-to-image model by updating model parameters with a unique identifier and a few reference images similar to DreamBooth (Ruiz et al., 2023). After personalization, we employ RL fine-tuning to address potential overfitting issues. During RL fine-tuning, the personalized model generates new images of the subject using prompts designed to ensure the model reflects the desired characteristics faithfully. The model then receives rewards for its outputs based on the image-text alignment. We update the model to maximize the rewards using a policy gradient method (Black et al., 2023; Fan et al., 2023). This iterative process effectively mitigates overfitting and enables the model to generate subject images with a high level of alignment to the provided text descriptions (see Figure 3). Furthermore, we introduce complementary techniques to enhance the synergy between the two processes. By combining all proposed techniques, our method can generate personalized images of the Paris 2024 Olympic mascot plushie that align with text prompts (see Figure 1)

We evaluate the performance of our proposed method on a diverse set of subject images, comparing it to prior methods, such as DreamBooth (Ruiz et al., 2023), Custom Diffusion (Kumari et al., 2023), NeTI (Alaluf et al., 2023), and Textual Inversion (Gal et al., 2022). In both quantitative and qualitative evaluations, our method outperforms the other methods in terms of image-text alignment, while maintaining high subject fidelity. Additionally, our human evaluation shows that human raters prefer InstructBooth over other models in side-by-side comparisons.

2. InstructBooth

2.1. Personalizing Text-to-Image Models

Prompts with Unique Identifiers for Personalization. To generate new images of a specific subject, given only a few reference images, we leverage DreamBooth (Ruiz et al., 2023). DreamBooth fine-tunes a text-to-image diffusion model by associating each user-provided subject with a

unique identifier. Formally, given parameterized denoising function ϵ_θ, we fine-tune the diffusion model by minimizing the following loss:

$$\mathcal{L}_P := \mathbb{E}_{\mathbf{z}, \mathbf{z}^{pr}, \mathbf{c}, \mathbf{c}^{pr}, \epsilon, \epsilon', t, t'} \Big[\left\| \epsilon - \epsilon_\theta \left(\mathbf{z}_t, t, \mathbf{c} \right) \right\|_2^2$$
$$+ \left\| \epsilon' - \epsilon_\theta \left(\mathbf{z}_{t'}^{pr}, t', \mathbf{c}^{pr} \right) \right\|_2^2 \Big],$$

where ϵ, ϵ' are Gaussian noise, $\mathbf{c}$ is a prompt with a unique identifier (i.e., "a [identifier] [class noun]"), $\mathbf{c}^{pr}$ is a prompt consisting of only a class noun (i.e., "a [class noun]"), $\mathbf{z}_t$ is a noised latent representation of the specific subject image at timestep t, and $\mathbf{z}_{t'}^{pr}$ is a noised latent representation of the image representing the class to which a specific subject belongs at timestep t'. The first term represents personalization loss, where the model learns a given subject with a unique identifier (i.e., "[identifier]"), and the second term is prior-preservation loss, which is used to supervise the model with its own generated images, to retain its visual prior. This fine-tuning process enables the model to generate new images of a subject in different contexts while maintaining its visual prior, which is crucial in performance.

Detailed Descriptions for Rare Subjects. In many cases, when fine-tuning a text-to-image model using a unique identifier with a class noun (i.e., "a [identifier] [class noun]"), it leads to the generation of perceptually similar personalized images. However, we found that the model often struggles to learn a rare subject with unseen visual attributes. For example, when personalizing text-to-image models with the prompt "a [identifier] plushie", the model often fails to generate Phryge (the Paris 2024 Olympic mascot) as a personalized output (see Figure 15 for supporting experimental results). This highlights the challenges posed by rare subjects with unseen visual attributes. We assume that using a class noun alone may not be sufficient to guide the text-to-image model to learn the distinctive characteristics of such a rare subject. Therefore, we add a detailed description of the subject's attribute to the prompt as the format of "a [identifier] [description] [class noun]". We observe that such a simple technique enables text-to-image models to capture the visual characteristics more concretely, resulting in improved personalization capabilities.

2.2. RL Fine-tuning for Improving Alignment

Fine-tuning a text-to-image model with a unique identifier enables the model to generate new images that are perceptually similar to a user-provided subject. However, such a personalized model often exhibits low text fidelity, especially in terms of contextual diversity, such as subjects' poses, articulations, or interactions with other objects. To address this issue, we propose fine-tuning the personalized model using reinforcement learning (RL). During the RL fine-tuning step, the personalized model is trained to max-
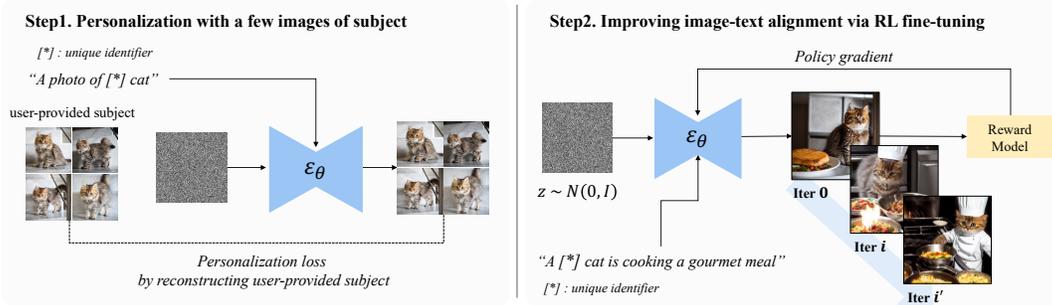
Figure 2. **An overview of InstructBooth**. Our method consists of two main steps: (left) Personalization with a few images of subject, where a pre-trained text-to-image model is fine-tuned with a unique identifier and (right) RL fine-tuning for improving image-text alignment, where we further fine-tune the personalized model to maximize the reward that quantifies image-text alignment.

imize a reward function reflecting image-text alignment. This training helps the model generate subject images that are closely aligned with the provided prompts.

Formally, similar to existing approaches (Black et al., 2023; Fan et al., 2023), we formulate the denoising process as multi-step MDP and treat transition distribution $p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{c})$ (described in Equation 4) as a policy. Based on this formulation, we fine-tune the personalized text-to-image model to maximize the expected reward $r(\mathbf{z}_0, \mathbf{c})$ of the generated images given the prompt distribution $p(\mathbf{c})$:

$$\max_\theta \quad \mathbb{E}_{p(\mathbf{c}), \, p_\theta(\mathbf{z}_0|\mathbf{c})} \left[r(\mathbf{z}_0, \mathbf{c}) \right], \qquad (1)$$

where $p_\theta(\cdot|\mathbf{c})$ is the sample distribution for the final denoised image $\mathbf{z}_0$. The gradient of the RL objective in Equation 1 can be rewritten as follows:

$$\mathbb{E}_{p(\mathbf{c}), p_\theta(\mathbf{z}_0|\mathbf{c})} \left[\sum_{t=1}^{T} \nabla_\theta \log p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{c}) \, r(\mathbf{z}_0, \mathbf{c}) \right], \quad (2)$$

where the proof is in Fan et al. (2023). Given the gradient of the RL objective, we update the model parameters θ to maximize the expected reward. To measure the alignment of generated images with text prompts, we use ImageReward (Xu et al., 2023), an open-source reward model trained on a large human feedback dataset. Xu et al. (2023) demonstrated ImageReward has a better correlation with human judgments compared to other scoring functions, such as CLIP (Radford et al., 2021) and BLIP (Li et al., 2022).

Text prompts for RL fine-tuning. For RL fine-tuning, we require training text prompts. One of the challenges we encounter when dealing with personalized models is the generation of images depicting personalized subjects in various poses. To address this challenge, we propose a templated approach that combines a class noun with a unique identifier along with the phrase describing a specific pose or activity. For example, we use prompts like "A [identifier] cat is cooking a gourmet meal" in Figure 2. However, we find that only utilizing prompts with unique identifiers can result in slow

RL fine-tuning, as the overfitted model merely generates good samples to get high reward signals. To mitigate this issue, we also utilize prompts without unique identifiers. In summary, we employ both types of prompts during RL fine-tuning: (i) "a [identifier] [class noun] [activity descriptor]" and (ii) "a [class noun] [activity descriptor]".

3. Experiments

3.1. Dataset

To evaluate the performance of personalized text-to-image models, we employ the DreamBench dataset (Ruiz et al., 2023), which consists of 30 unique subjects including various objects, live subjects, and pets. We evaluate the models using 25 diverse prompts, encompassing recontextualization, property modification, and accessorization. However, text prompts in DreamBench are not designed to evaluate the model's ability to generate images with actions (e.g., "a [identifier] [class noun] is cooking a gourmet meal") or interactions with other objects (e.g., "a [identifier] [class noun] with popcorn in it"). To address this, we introduce a new dataset comprising eight subjects and five phrases describing activities or interactions . Specifically, we use images of subjects sampled from prior works (Ruiz et al., 2023; Kumari et al., 2023) and generate a total of 40 prompts by combining subjects and phrases.

3.2. Qualitative Analysis

We compare the quality of personalized text-to-image generation against four baselines: DreamBooth (Ruiz et al., 2023), Custom diffusion (Kumari et al., 2023), NeTI (Alaluf et al., 2023), and Textual Inversion (Gal et al., 2022). As shown in Figure 3, DreamBooth produces images of visually similar subjects but often fails to accurately represent the context from the text prompt. For example, activities like "cooking" and "playing" are often not reflected in their generated images. Custom Diffusion generates images that align with the text prompt, but it exhibits low similarity to the provided subject. In contrast to baselines, InstructBooth generates
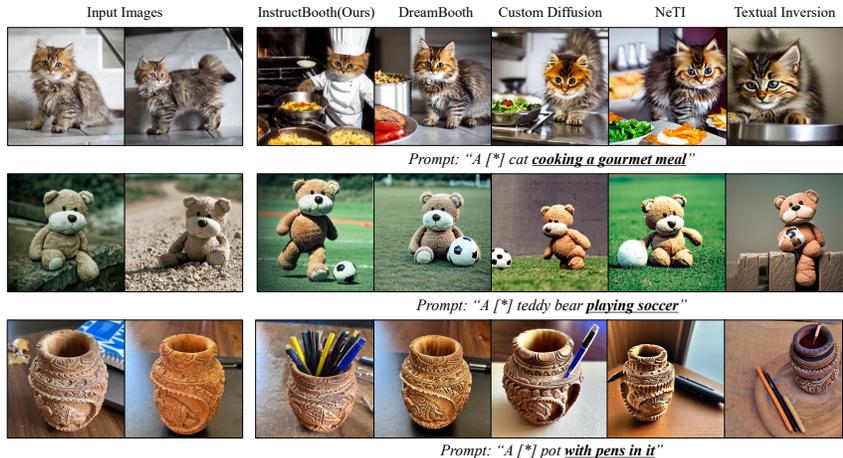
Figure 3. Qualitative comparison against alternative approaches. Given a few images of a unique subject (e.g., cat, teddy bear, and pot) and a text prompt, models are required to generate personalized images that align with the prompt. [*] denotes a unique identifier.
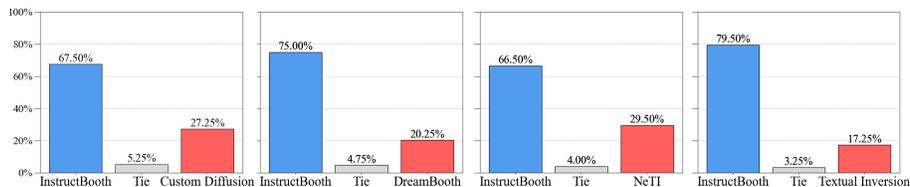

Figure 4. Human evaluation results between InstructBooth and baselines. Given two images generated by each model, we ask human raters to indicate which is better in overall quality.

images with high text alignment without sacrificing subject similarity. Moreover, our method effectively produces personalized images that closely match the intended context, including variations in subject poses (e.g., a teddy bear dribbling the ball) and costumes (e.g., a cat wearing like a chef).

3.3. Human Evaluation

To evaluate the quality of personalized text-to-image generation, we conduct a human study comparing InstructBooth with baselines (We provide more details of human evaluation in Appendix F). Table 1 summarizes the binary feedback on subject fidelity and text fidelity. The results show that the performance of baselines is limited due to a trade-off between subject fidelity and text fidelity. In contrast, InstructBooth outperforms baselines in terms of text fidelity while exhibiting superior perceptual similarity of subjects. Figure 4 shows the human preference rating in terms of overall quality. Due to improved text fidelity without compromising on subject fidelity, the images generated by InstructBooth are preferred at least twice as much as baselines when considering all relevant factors.

4. Conclusion

In this paper, we propose InstructBooth, a new method for improving the image-text alignment of personalized text-to-image models. We demonstrate that fine-tuning the model

Table 1. Human evaluation results.

Method	Text Fidelity	Subject Fidelity
Custom Diffusion	91.5%	92.8%
DreamBooth	81.3%	**97.0%**
NeTI	85.5%	93.8%
Textual Inversion	60.0%	80.0%
InstructBooth (Ours)	**97.3%**	**97.0%**

with RL in a subsequent stage mitigates overfitting, enabling the model to generate images of specific subjects with contextual diversity in terms of poses and interactions. In qualitative comparison and human evaluation, we show that InstructBooth can generate images that are more aligned with human preferences than those of existing models. We believe that our approach of subsequent fine-tuning broadens the potential of personalized text-to-image models by allowing the usage of diverse prompts.

Acknowledgements

References

Agarwal, A., Karanam, S., Shukla, T., and Srinivasan, B. V. An image is worth multiple words: Multi-attribute inversion for constrained text-to-image synthesis. *arXiv preprint arXiv:2311.11919*, 2023.

Alaluf, Y., Richardson, E., Metzer, G., and Cohen-Or, D. A neural space-time representation for text-to-image personalization. *arXiv preprint arXiv:2305.15391*, 2023.

Black, K., Janner, M., Du, Y., Kostrikov, I., and Levine, S. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *IEEE/CVF international conference on computer vision*, 2021.

Chang, H., Zhang, H., Barber, J., Maschinot, A., Lezama, J., Jiang, L., Yang, M.-H., Murphy, K., Freeman, W. T., Rubinstein, M., et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023.

Clark, K., Vicol, P., Swersky, K., and Fleet, D. J. Directly fine-tuning diffusion models on differentiable rewards. *arXiv preprint arXiv:2309.17400*, 2023.

Fan, Y., Watkins, O., Du, Y., Liu, H., Ryu, M., Boutilier, C., Abbeel, P., Ghavamzadeh, M., Lee, K., and Lee, K. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. In *Advances in neural information processing systems*, 2023.

Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., and Cohen-Or, D. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

Han, L., Li, Y., Zhang, H., Milanfar, P., Metaxas, D., and Yang, F. Svdiff: Compact parameter space for diffusion fine-tuning. *arXiv preprint arXiv:2303.11305*, 2023.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In *Advances in neural information processing systems*, 2020.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., and Levy, O. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *Advances in neural information processing systems*, 2023.

Kumari, N., Zhang, B., Zhang, R., Shechtman, E., and Zhu, J.-Y. Multi-concept customization of text-to-image diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Lee, K., Liu, H., Ryu, M., Watkins, O., Du, Y., Boutilier, C., Abbeel, P., Ghavamzadeh, M., and Gu, S. S. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023.

Li, J., Li, D., Xiong, C., and Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 2022.

Liu, Z., Feng, R., Zhu, K., Zhang, Y., Zheng, K., Liu, Y., Zhao, D., Zhou, J., and Cao, Y. Cones: Concept neurons in diffusion models for customized generation. *arXiv preprint arXiv:2303.05125*, 2023.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *ICLR*, 2019.

Prabhudesai, M., Goyal, A., Pathak, D., and Fragkiadaki, K. Aligning text-to-image diffusion models with reward backpropagation. *arXiv preprint arXiv:2310.03739*, 2023.

Qiu, Z., Liu, W., Feng, H., Xue, Y., Feng, Y., Liu, Z., Zhang, D., Weller, A., and Schölkopf, B. Controlling text-to-image diffusion by orthogonal finetuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=K30wTdIIYc.

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*, 2021.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, 2015.

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

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. In *Advances in Neural Information Processing Systems*, 2022.

Tewel, Y., Gal, R., Chechik, G., and Atzmon, Y. Key-locked rank one editing for text-to-image personalization. In *ACM SIGGRAPH 2023 Conference Proceedings*, 2023.

Voynov, A., Chu, Q., Cohen-Or, D., and Aberman, K. $p+$: Extended textual conditioning in text-to-image generation. *arXiv preprint arXiv:2303.09522*, 2023.

Wu, X., Sun, K., Zhu, F., Zhao, R., and Li, H. Human preference score: Better aligning text-to-image models with human preference. In *IEEE/CVF International Conference on Computer Vision*, 2023.

Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., and Dong, Y. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *Advances in neural information processing systems*, 2023.

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*, 2022.

Zhang, Y., Dong, W., Tang, F., Huang, N., Huang, H., Ma, C., Lee, T.-Y., Deussen, O., and Xu, C. Prospect: Expanded conditioning for the personalization of attribute-aware image generation. *arXiv preprint arXiv:2305.16225*, 2023.

A. Related Work

Personalized Text-to-Image Generation. Since text-to-image models (Ramesh et al., 2022; Saharia et al., 2022; Rombach et al., 2022; Chang et al., 2023; Yu et al., 2022) have shown impressive results, several studies have been proposed to personalize the text-to-image models using only a few images of a specific subject. Specifically, two distinct streams of research have emerged to achieve this goal. The first stream, based on Textual Inversion (Gal et al., 2022), focuses on optimizing new word embedding to represent a given subject or concept. This line of research demonstrates high controllability and has recently been expanded to learn new word embedding in various embedding spaces (Alaluf et al., 2023; Zhang et al., 2023; Voynov et al., 2023; Agarwal et al., 2023). The second stream, based on DreamBooth (Ruiz et al., 2023), involves the method to fine-tune the pre-trained models using the text prompt with a unique identifier. Within this line of research, each approach is differentiated by determining the scope of weights to be learned, such as fine-tuning only the weights of the attention layer (Kumari et al., 2023; Tewel et al., 2023; Han et al., 2023; Qiu et al., 2023). Our method is based on the latter (i.e., DreamBooth) in terms of personalizing the pre-trained models using a unique identifier. However, unlike prior works, we introduce reinforcement learning (RL) in a subsequent step. This incorporation of RL effectively mitigates overfitting, enabling the generation of images with high text fidelity for prompts that have been challenging for existing methods.

Improving Image-Text Alignment. Despite the impressive success of text-to-image models, they often struggle to generate images that accurately align with text prompts. To address this issue, recent studies have investigated *learning from human feedback* in text-to-image generation (Lee et al., 2023; Xu et al., 2023; Kirstain et al., 2023; Wu et al., 2023). These methods first learn a reward function intended to reflect what humans care about in the task, using human feedback on model outputs. Subsequently, they utilize the learned rewards to enhance image-text alignment through techniques such as rejection sampling (Kirstain et al., 2023), reward-weighted learning (Lee et al., 2023; Wu et al., 2023), and direct reward optimization via gradient (Xu et al., 2023; Clark et al., 2023; Prabhudesai et al., 2023). DDPO (Black et al., 2023) and DPOK (Fan et al., 2023) formulate the fine-tuning problem as a multi-step decision-making problem and propose a policy gradient method to maximize the expected rewards. They demonstrate that RL fine-tuning can effectively improve text-to-image alignment. Inspired by their successes, we propose to utilize the fine-tuning with the reward model to improve the image-text alignment of personalized text-to-image models.

B. Preliminary

In this work, we consider text-to-image diffusion models, a class of generative models that transform Gaussian noise via iterative denoising process to model data distribution $p(\mathbf{x})$ (Ho et al., 2020). Specifically, we utilize a latent diffusion model (Rombach et al., 2022), which (i) operates in a highly compressed lower-dimensional latent space $\mathbf{z} = \mathcal{E}(\mathbf{x})$ using an image encoder $\mathcal{E}$, rather than relying on pixel-based image representations $\mathbf{x}$, and (ii) utilizes a conditional denoising autoencoder $\epsilon_\theta(\mathbf{z}, t, \mathbf{c})$ to control the synthesis process with inputs $\mathbf{c}$ (i.e., language prompts) by modeling conditional distributions $p(\mathbf{z}|\mathbf{c})$. The training of this model involves predicting the added noise ϵ to the latent representation $\mathbf{z}_t$ at timestep t, using an additional condition $\mathbf{c}$. Formally, the objective of the training is as follows:

$$\mathcal{L}_{LDM} := \mathbb{E}_{\mathbf{z}, \mathbf{c}, \epsilon, t} \left[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \mathbf{c})\|_2^2 \right], \tag{3}$$

where t is timestep uniformly sampled from $\{1, 2, \ldots, T\}$, ϵ is a Gaussian noise $\sim \mathcal{N}(0, I)$, and $\mathbf{z}_t$ is the noised latent representation at a timestep t. At inference phase, the iterative denoising process is conducted to produce the denoised sample $\mathbf{z}_0$, using the noise predicted by $\epsilon_\theta(\mathbf{z}_t, t, \mathbf{c})$. Specifically, the predicted noise is used to calculate the mean of transition distribution $p_\theta(\mathbf{z}_{t-1} \mid \mathbf{z}_t, \mathbf{c})$ in denoising process. The formulation is as follows:

$$\mu_\theta(\mathbf{z}_t, t, \mathbf{c}) = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{z}_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(\mathbf{z}_t, t, \mathbf{c}) \right), p_\theta(\mathbf{z}_{t-1} \mid \mathbf{z}_t, \mathbf{c}) = \mathcal{N}(\mu_\theta(\mathbf{z}_t, t, \mathbf{c}), \Sigma_t), \tag{4}$$

where α_t, β_t are pre-defined constants for timestep dependant denoising, Σ_t is covariance matrix of denoising transition, and $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$. Lastly, the decoder $\mathcal{D}$ transforms the denoised latent sample $\mathbf{z}_0$ into $\mathbf{x}$ (i.e., pixel space). In the following sections, we treat $\mathbf{z}_0$ as final image, omitting the decoder process for brevity.

C. Setup

Dataset Details   As we explained in the main paper, the current DreamBench (Ruiz et al., 2023) dataset is not well-suited to evaluate the model's ability to generate images regarding actions or interactions with other objects. Thus, as shown in Figure 5, we introduce a new dataset, which consists of eight subjects (i.e., cat, dog, teddy bear, monster toy, wooden pot, cup, motorbike, and bike) and five corresponding phrases describing activities (e.g., "playing soccer," "riding a bike," and "cooking") and interactions (e.g., "floating on the water" and "with popcorn in it"). Overall, we have $40 (= 8 \times 5)$ object-phrase pairs. Note that we use images of a dog and a monster toy from Ruiz et al. (2023) and images of a cat, a teddy bear, a wooden pot, a cup, a motorbike, and a bike from Kumari et al. (2023). Further, we group two similar subjects (e.g., a cat and a dog) and use the same five phrases for each group.

Subject Images		Text Prompts
Cat	Dog	• "A [identifier] [class noun] is cooking a gourmet meal" • "A [identifier] [class noun] dressed in flamboyant attire dancing at a party venue" • "A [identifier] [class noun] acting in a play wearing a costume" • "A [identifier] [class noun] playing a guitar on the stage" • "A [identifier] [class noun] wearing a headphone and DJing at the club"
Teddy bear	Monster toy	• "A [identifier] [class noun] playing soccer" • "A [identifier] [class noun] playing tennis" • "A [identifier] [class noun] doing taekwondo" • "A [identifier] [class noun] playing ice hockey" • "A [identifier] [class noun] riding a bike"
Wooden pot	Cup	• "A [identifier] [class noun] with pens in it" • "A [identifier] [class noun] floating on the water" • "A [identifier] [class noun] on a flower garden with flowers in it" • "A cat is poking face out of a [identifier] [class noun]" • "A [identifier] [class noun] with popcorn in it"
Motorbike	Bike	• "A [identifier] [class noun] driving through a vibrant city at night" • "A [identifier] [class noun] cornering on the cliffside road" • "A rider on a [identifier] [class noun]" • "A [identifier] [class noun] racing in a competition" • "A [identifier] [class noun] driving on a highway"

Figure 5. Our newly created evaluation dataset (for text-to-image personalization methods) that contains phrases focusing on actions or interactions with other objects. Our dataset consists of eight subjects (i.e., cat, dog, teddy bear, monster toy, wooden pot, cup, motorbike, and bike) and five corresponding phrases describing activities (e.g., "playing soccer," "riding a bike," and "cooking") or interactions (e.g., "floating on the water" and "with popcorn in it").

Evaluation Metrics   To assess the quality of text-to-image personalization models, we evaluate them based on two key aspects: (i) text fidelity, which measures the alignment between the text prompt and the generated image, and (ii) subject fidelity, which measures the preservation of subject details in generated images. For text fidelity, we use three metrics: CLIP-T (Radford et al., 2021), ImageReward (Xu et al., 2023), and PickScore (Kirstain et al., 2023). CLIP-T (Radford et al., 2021) measures the cosine similarity between image and text embeddings. ImageReward and PickScore are scoring functions trained on a large human feedback dataset to measure image-text alignment. For subject fidelity, we use DINO score which measures the cosine similarity between the generated and reference images in ViTS/16 DINO (Caron et al., 2021) embedding spaces.

Implementation Details   As for personalization step, our fine-tuning pipeline is built upon the publicly available repository.[1] We use Stable Diffusion v1.5 (SD; (Rombach et al., 2022)) as our baseline text-to-image mode and generate 200 images representing the class to which input subject belongs for prior-preservation loss. We set the learning rate to 2×10^{-5} and batch size to 2 (one is an input subject image for personalization loss, and the other is a class image for prior-preservation loss). We fine-tune the entire U-Net (Ronneberger et al., 2015) using AdamW (Loshchilov & Hutter, 2019), where $\beta_1 = 0.9$, $\beta_2 = 0.99$ and weight decay 0.01.

As for RL fine-tuning step, our implementation is based on DDPO (Black et al., 2023). We generate 16 images at each

---

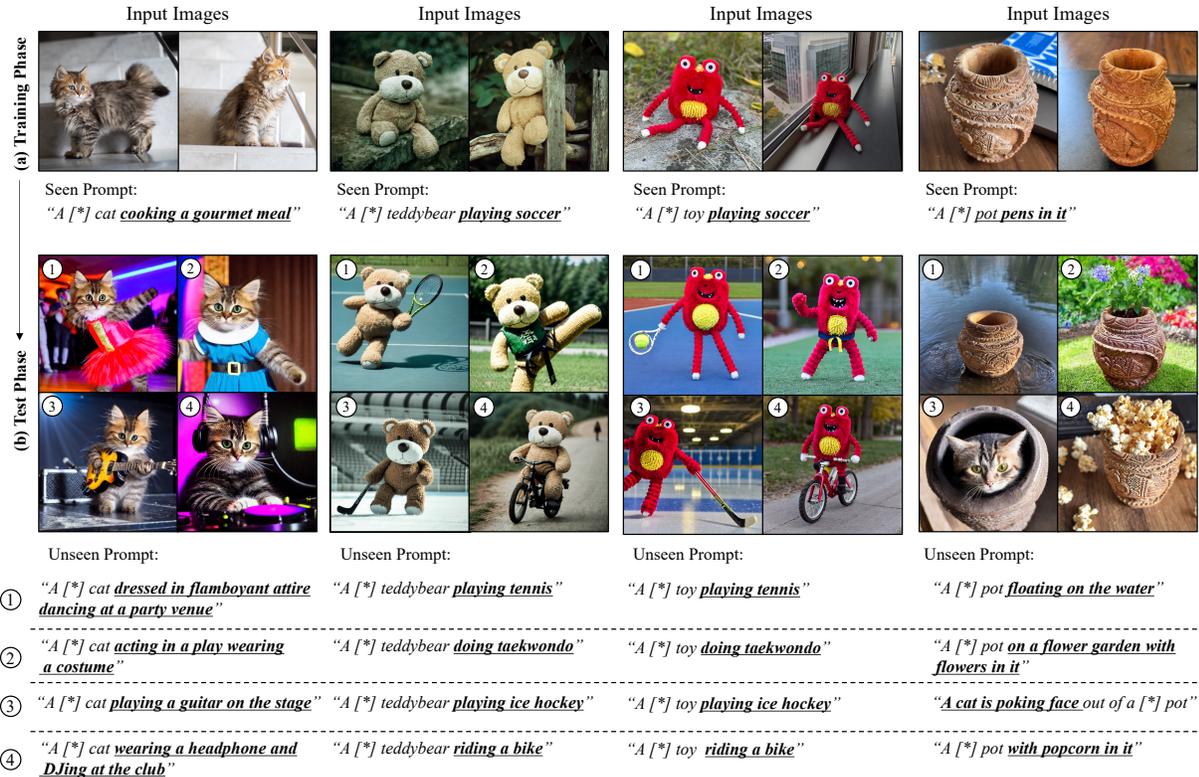[1]https://github.com/huggingface/diffusers/tree/main/examples/dreambooth

Figure 6. Samples generated by InstructBooth on unseen text prompts. Our method generates personalized images with high image-text alignment. [*] denotes a unique identifier.

epoch using a personalized diffusion model with an inference step of 50 and a guidance scale of 7.5. Among the denoising trajectories of 16 generated images, we randomly use 8 samples as the training batch at each gradient step. For policy gradient-based optimization, we use importance sampling with a clip range of 1×10^{-4}. We set the learning rate to 2×10^{-5} and update the model using AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.99$ and weight decay 0.01. Note that we apply Low-Rank Adaption (LoRA; (Hu et al., 2021)) to the U-Net in RL fine-tuning step.

D. Additional Experimental Results

Generalization to Unseen Prompts   To understand the generalization ability, we evaluate InstructBooth using unseen text prompts. Figure 6 shows image samples from InstructBooth on four different subjects with unseen prompts. Our model successfully generates images of user-provided subjects engaged in various unseen activities and costumes, demonstrating the generalization effects of RL fine-tuning. For example, when InstructBooth is trained with the prompt "cooking a gourmet meal", it can still generate images of the personalized cat engaged in unrelated activities. Our RL fine-tuning encourages the model to generate subjects in different actions or poses, preventing the model from simply copying and pasting the given subjects. Such a learning process helps avoid overfitting and enables the model to generate more diverse and text-aligned actions.

Quantitative Analysis   We conduct a quantitative evaluation by measuring several metrics introduced in Appendix C. For the evaluation, similar to a human evaluation, we use 40 prompts related to actions to create a total of 400 images, with 10 images for each prompt. As shown in Table 2, InstructBooth achieves the highest scores in all metrics related to text fidelity. Furthermore, unlike Custom Diffusion which excels in text fidelity but struggles with subject fidelity, our method maintains competitive subject fidelity. However, in contrast to human evaluation results, our method exhibits a slightly lower score in subject fidelity compared to DreamBooth. This difference can be attributed to the limitations of the subject fidelity metric, i.e., DINO, which penalizes changes in pose and primarily considers appearance resemblance. In our tested prompts that emphasize actions, even if the personalized image is successfully created, it may receive a lower DINO score. To provide further support for this explanation, we include the images and their corresponding scores in Appendix E.

Table 2. Comparison with Dreambooth, Custom Diffusion, NeTI and Textual Inversion on our dataset.

	Text Fidelity				Subject Fidelity	
Method	CLIP-T	ImageReward	PickScore	Human	DINO	Human
Custom Diffusion	**0.323**	1.088	0.224	91.5%	0.535	92.8%
DreamBooth	0.310	0.441	0.220	81.3%	**0.699**	**97.0%**
NeTI	0.311	0.480	0.218	85.5%	0.643	93.8%
Textual Inversion	0.271	-1.15	0.206	60.0%	0.576	80.0%
InstructBooth (Ours)	**0.323**	**1.196**	**0.227**	**97.3%**	0.650	**97.0%**

Table 3. Comparison with baselines on DreamBench. For baselines, we report both performances reported in Ruiz et al. (2023) and those obtained through our implementation (denoted as our impl).

	Text Fidelity			Subject Fidelity
Method	CLIP-T	ImageReward	PickScore	DINO
Textual Inversion	0.255	N/A	N/A	0.569
Textual Inversion (our impl)	0.257	-1.337	0.203	0.537
DreamBooth	0.305	N/A	N/A	**0.668**
DreamBooth (our impl)	0.297	0.052	0.214	0.648
InstructBooth (Ours)	**0.306**	**0.651**	**0.217**	0.661

DreamBench Results   In addition to the experiments with our dataset, we also evaluate InstructBooth with Dream-Bench (Ruiz et al., 2023) dataset to demonstrate that our method can enhance text fidelity in diverse prompt scenarios. Following DreamBench, we generate four images for each subject and each prompt, resulting in a total of 3,000 images for measurement. As for our prompts for RL fine-tuning, we set the two training prompts regarding recontextualization and accessorization which DreamBench primarily deals with. Note that we do not use DreamBench's prompts directly as training prompts, which means that all evaluation is performed with unseen prompts to our method. As shown in Table 3, InstructBooth achieves the highest scores in text fidelity, while also demonstrating competitive performances in terms of subject fidelity.

Additional Generated Examples   We provide additional generated examples comparing other existing approaches: Dream-Booth (Ruiz et al., 2023), Custom Diffusion (Kumari et al., 2023), NeTI (Alaluf et al., 2023), and Textual Inversion (Gal et al., 2022). Figure 7 shows the generated samples using *seen* prompt during RL fine-tuning. We provide additional examples with *unseen* prompts in Figure 8 and 9. Additionally, we provide the generated samples using DreamBench dataset in Figure 10.
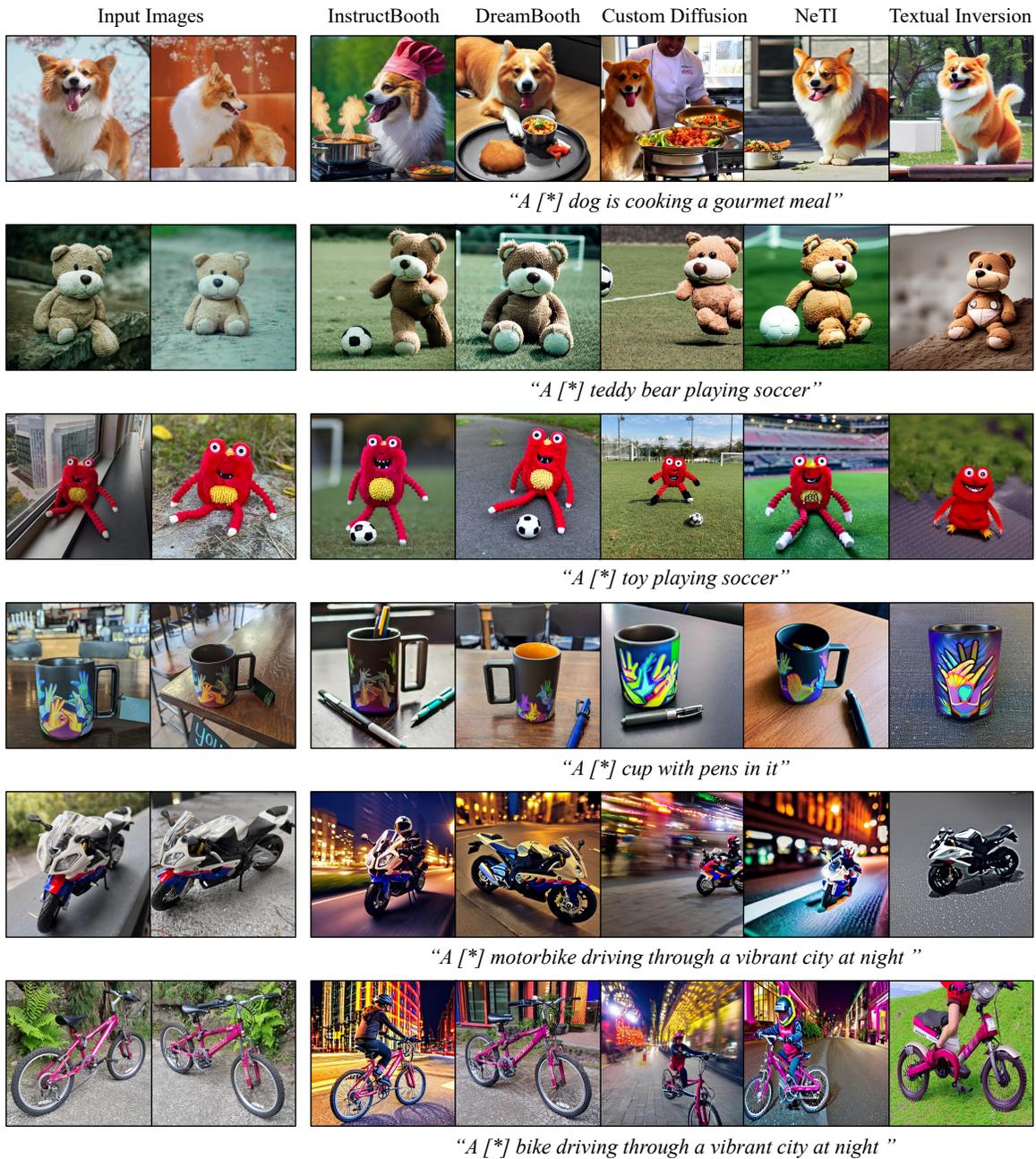
Figure 7. Additional examples generated by InstructBooth, DreamBooth (Ruiz et al., 2023), Custom Diffusion (Kumari et al., 2023), NeTI (Alaluf et al., 2023), and Textual Inversion (Gal et al., 2022). Note that the prompts used to generate these samples are used in the RL fine-tuning step of InstructBooth (i.e., *seen* prompt).
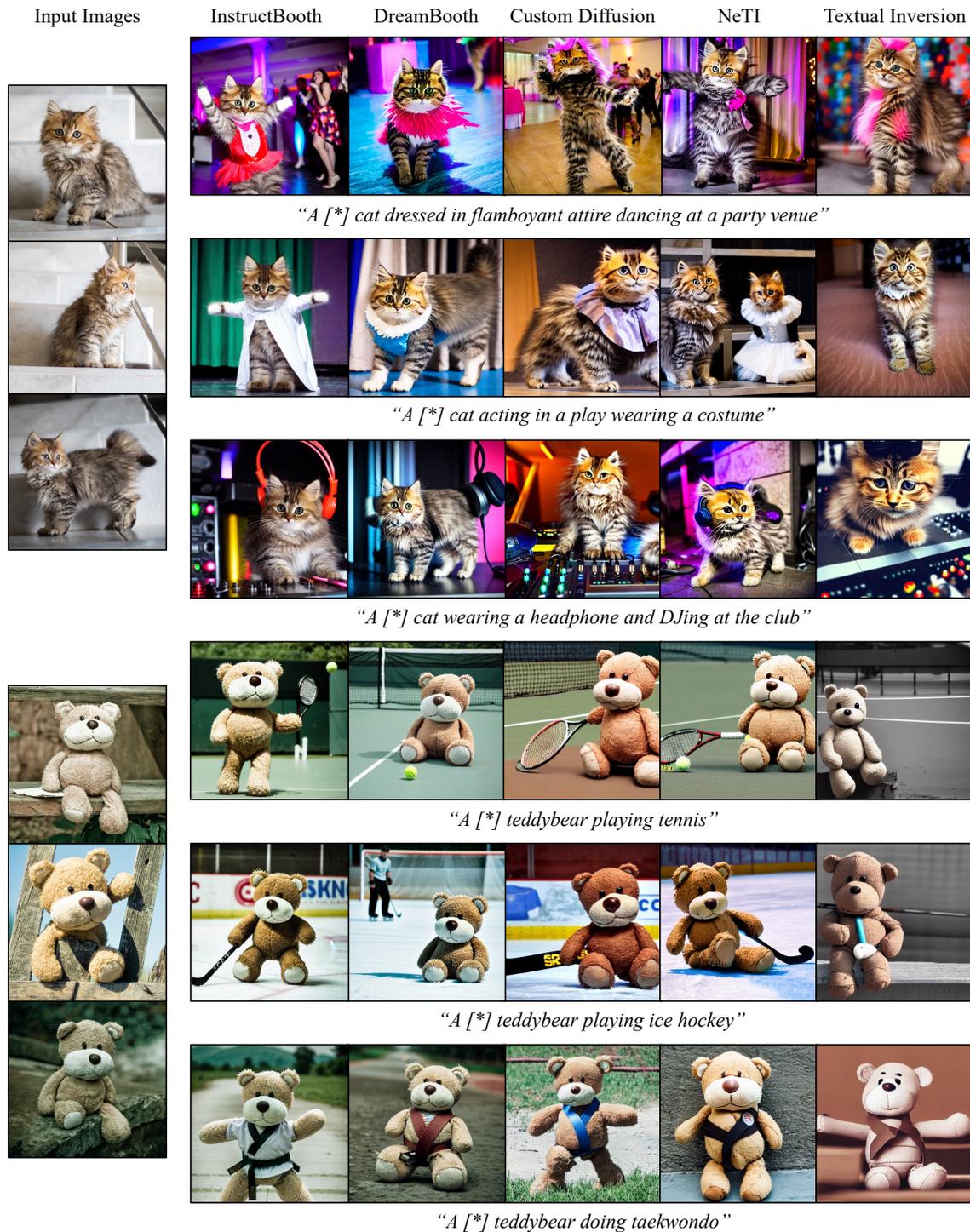
Input Images	InstructBooth	DreamBooth	Custom Diffusion	NeTI	Textual Inversion

"A [*] cat dressed in flamboyant attire dancing at a party venue"

"A [*] cat acting in a play wearing a costume"

"A [*] cat wearing a headphone and DJing at the club"

"A [*] teddybear playing tennis"

"A [*] teddybear playing ice hockey"

"A [*] teddybear doing taekwondo"

Figure 8. Additional examples generated by InstructBooth, DreamBooth (Ruiz et al., 2023), Custom Diffusion (Kumari et al., 2023), NeTI (Alaluf et al., 2023), and Textual Inversion (Gal et al., 2022). Note that the prompts used to generate these samples are not used in the RL fine-tuning step of InstructBooth (i.e., *unseen* prompt).
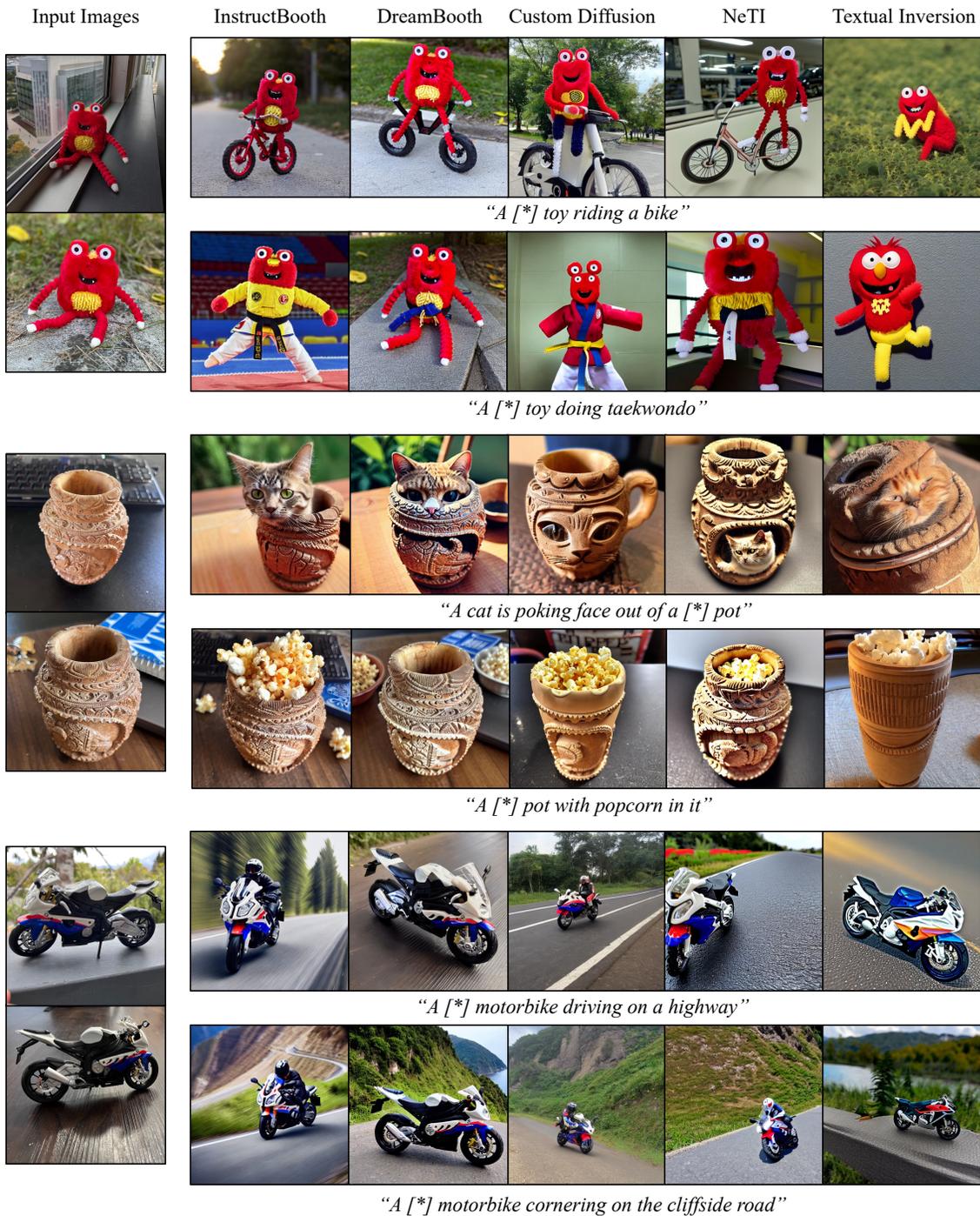
Input Images	InstructBooth	DreamBooth	Custom Diffusion	NeTI	Textual Inversion

"A [*] toy riding a bike"

"A [*] toy doing taekwondo"

"A cat is poking face out of a [*] pot"

"A [*] pot with popcorn in it"

"A [*] motorbike driving on a highway"

"A [*] motorbike cornering on the cliffside road"

Figure 9. Additional examples generated by InstructBooth, DreamBooth (Ruiz et al., 2023), Custom Diffusion (Kumari et al., 2023), NeTI (Alaluf et al., 2023), and Textual Inversion (Gal et al., 2022). Note that the prompts used to generate these samples are not used in the RL fine-tuning step of InstructBooth (i.e., *unseen* prompt).
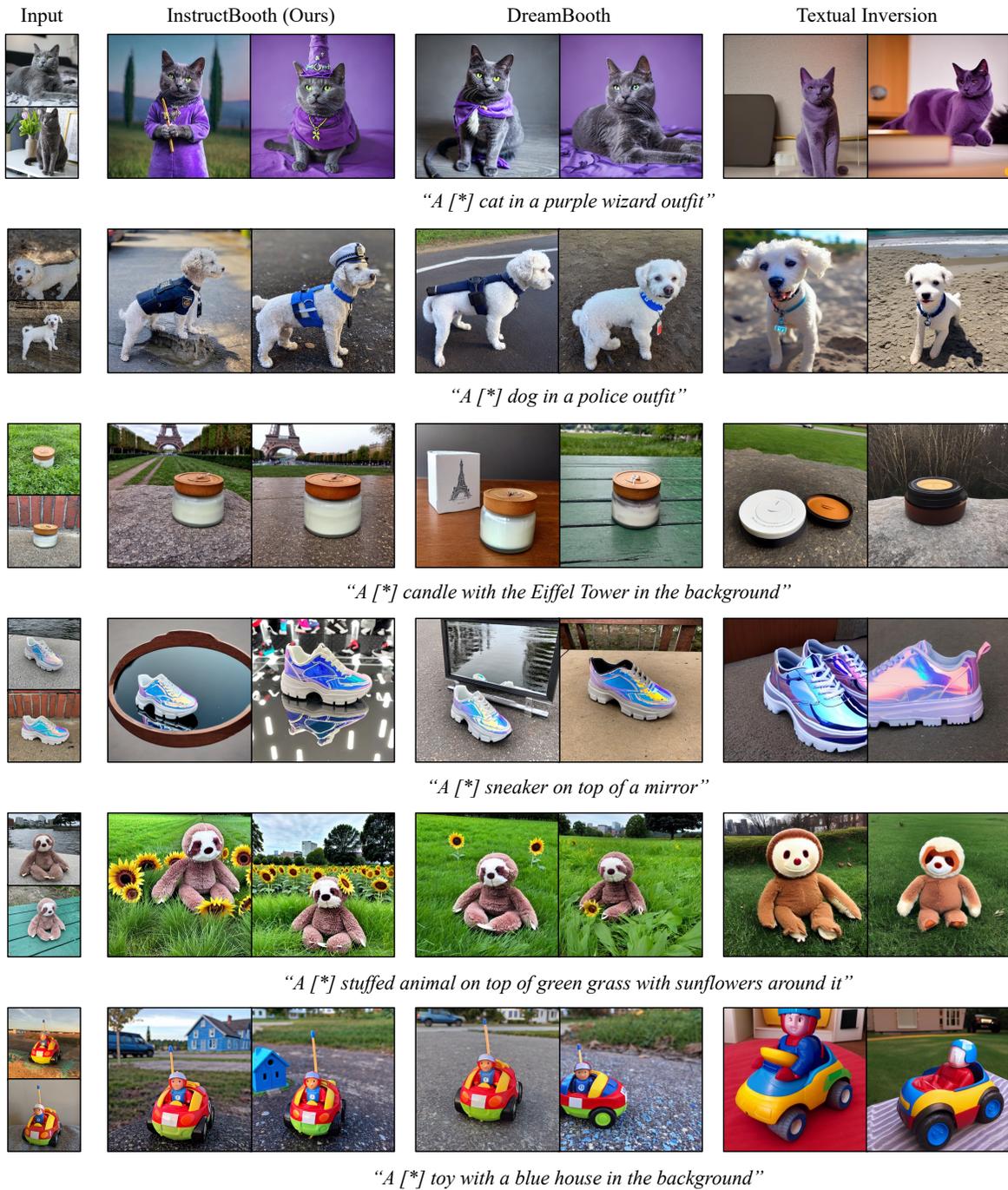
Figure 10. Samples generated by InstructBooth, DreamBooth (Ruiz et al., 2023) and Textual Inversion (Gal et al., 2022) using Dream-Bench's prompts.

E. Analysis of Subject Fidelity Metric

In the main paper, we argue that subject fidelity metric (i.e., DINO (Caron et al., 2021)) might not be a perfect metric to evaluate the quality of generated personalized images. In our analysis, this metric gives the best score for copied-and-pasted subjects, giving lower scores for the same subjects of different poses or actions. To demonstrate it, we present the generated results of DreamBooth (Ruiz et al., 2023) and InstructBooth along with the corresponding scores. For this analysis, we use "A [identifier] teddybear playing tennis" as a text prompt and collect 7 images of each method. Note that we intentionally collect DreamBooth's results containing unnatural visual elements to show that changes in pose incur a greater penalty than failure factors of personalization. As shown in Figure 11, even though examples from DreamBooth are perceptually *less* similar to the reference subject (e.g., a teddy bear with three legs), their DINO scores are clearly shown better than ours, where we generate perceptually *more* similar subjects with different actions. This may indicate that such subject fidelity metric may not be useful to accurately quantify perceptual similarities of subjects with different poses.

Input Image	Images generated by <u>DreamBooth</u>						
DINO(↑)	0.73	0.70	0.74	0.73	0.71	0.70	0.74
Average score : **0.72**							

Input Image	Images generated by <u>InstructBooth</u>						
DINO(↑)	0.51	0.57	0.60	0.66	0.68	0.63	0.64
Average score : **0.61**							

Figure 11. Examples of generated personalized images by DreamBooth (Ruiz et al., 2023) and InstructBooth (ours), using input image (left). We report subject fidelity scores (DINO (Caron et al., 2021)) for each generated image. Note that ↑ indicates that the higher number is the better.

F. Human Evaluation Details

Our human evaluation consists of three different parts: (i) subject fidelity evaluation, (ii) text fidelity evaluation, (iii) and overall quality evaluation. In (i), as shown in Figure 12, human raters are asked to answer yes or no to the following question: "Do the objects in the Image A closely resemble those in Given Image?" Importantly, human raters are asked to focus on perceptual resemblances without consideration of subjects' poses. We provide an example question on the instruction page to guide human raters to focus on perceptual resemblances.

Further, we evaluate text fidelity, as shown in Figure 13, by asking human raters to answer the following question: "Is the alignment of the image with the text correct?" Like previous subject fidelity evaluations, we provide instructions to guide human raters to focus on image-text alignments. Lastly, we also conduct overall quality evaluation by asking human raters to answer the following question: "When considered comprehensively from three different perspectives, please indicate which of the two images you prefer." In this question, we guide human raters to consider the following three perspectives with priority in order: text fidelity, subject fidelity, and naturalness (see Figure 14).
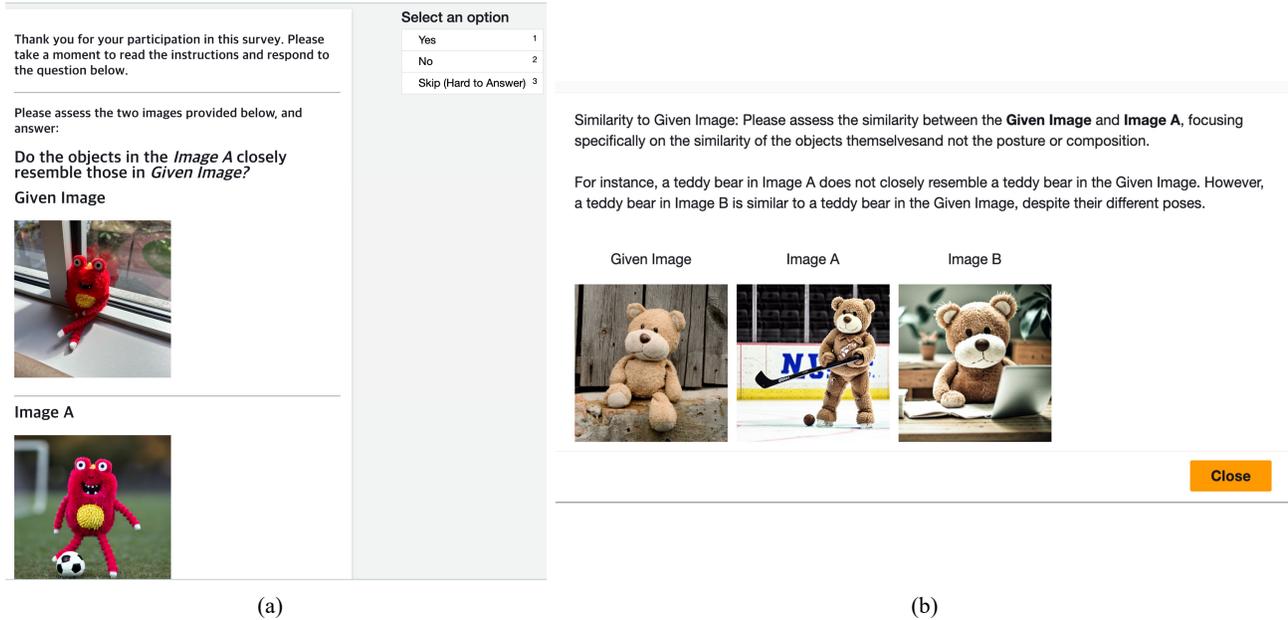
(a)

(b)

Figure 12. (a) A screenshot of our human evaluation questionnaires to evaluate subject fidelity with (b) an instruction.
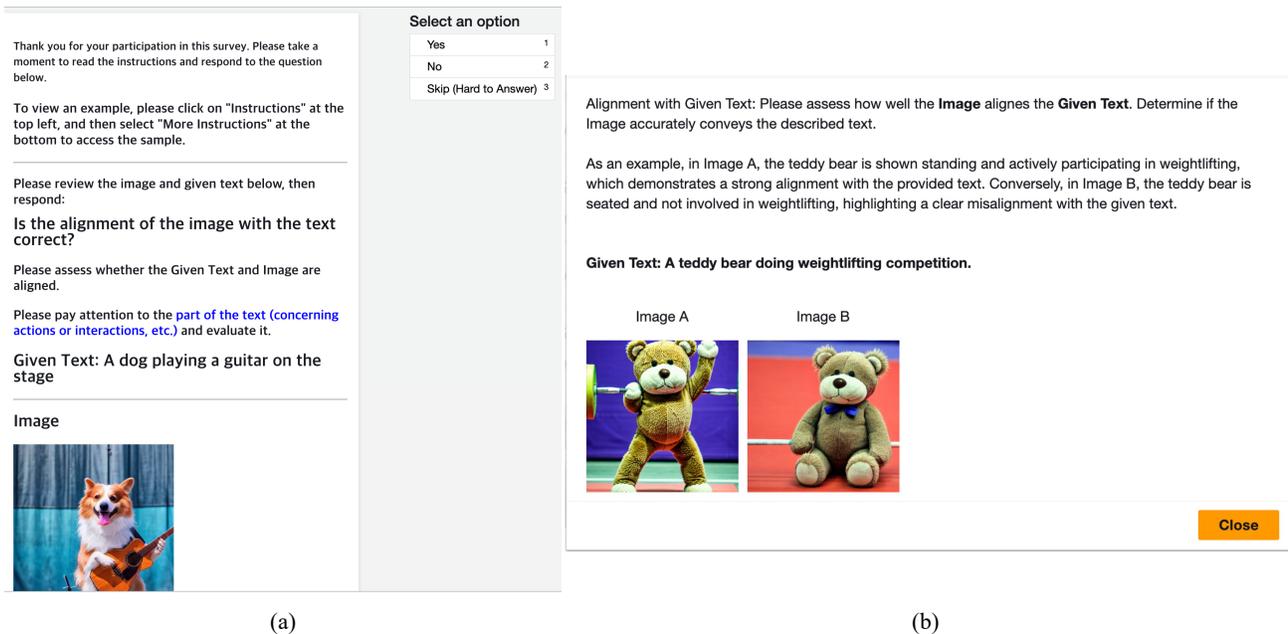

(a)

(b)

Figure 13. (a) A screenshot of our human evaluation questionnaires to evaluate text fidelity with (b) an instruction.

Thank you for your participation in this survey. Please take a moment to read the instructions and respond to the question below.

To view an example, please click on "Instructions" at the top left, and then select "More Instructions" at the bottom to access the sample.

Please examine the three images and the given text, and provide your judgments:

When considered comprehensively from three different perspectives, please indicate which of the two images you prefer.

The priority of judgment is 1-3.

1) alignment with the text
2) similarity to the object that exists in the Given Image
3) naturalness (consider image quality, recognizability, etc.)

Please pay attention to the part of the text (concerning actions or interactions, etc.) and evaluate it.

Given Text: A cat dressed in flamboyant attire dancing at a party venue

Given Image

Image A

Image B

Select an option

A	1
B	2
Can not Determine	3

A/B Test: Considering a thorough assessment based on three distinct perspectives (1. alignment with the text, 2. similarity with the given image, 3. naturalness), kindly indicate your preference between the two images.

For instance, Image A exhibits a strong alignment with the text, featuring a teddy bear holding a hockey club in the background of an ice rink, and closely resembling the teddy bear in the Given Image. In contrast, Image B lacks the presence of clothing or a hockey club to suggest hockey play, despite the ice rink in the background. Additionally, the teddy bear's disproportionately large eyes do not closely resemble the teddy bear in the Given Image. Therefore, Image A is the preferred choice.

Given Image   Image A   Image B

Close

(a)   (b)

Figure 14. (a) A screenshot of our human evaluation questionnaires to evaluate overall quality with (b) an instruction.

G. Ablation study

Effects of Detailed Descriptions in Personalization.
To demonstrate the importance of detailed description when personalizing text-to-image models with a rare subject, we conduct an ablation study on personalization techniques. In this study, we use Phryge (the Paris 2024 Olympic mascot) as our target subject. Since this mascot was not released when our base text-to-image model (i.e., Stable Diffusion v1.5) was trained, it can be considered a truly rare concept for the text-to-image model to learn. We compare the model trained with a detailed prompt "a
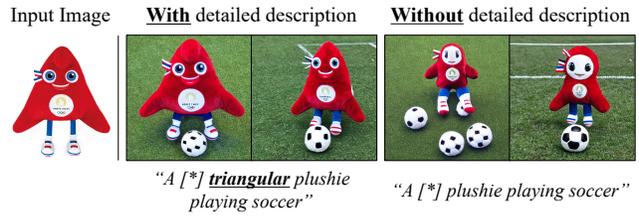

Figure 15. Comparison of results generated by models trained with and without prompts including a detailed description (i.e., **triangular**).[*] denotes a unique identifier.

[*] triangular plushie" and that trained with a standard simple prompt "a [*] plushie" using images of Phryge. As shown in Figure 15, the model trained with a standard simple prompt fails to accurately represent the subject's features.

Text Prompts in RL Fine-tuning.   In Section 2.2, we introduce an important trick to leverage both text prompts, with and without a unique identifier for RL fine-tuning. To verify its effectiveness, we compare two models: one trained using a prompt with a unique identifier and another trained using both prompts, with and without a unique identifier. In this comparison, we use "Wooden pot" and "Cup" as the target subject, and set "with the pens in it" as the target prompt. As shown in Figure 16, the model trained with both prompts exhibits improved sample efficiency, demonstrating that our proposed technique mitigates the challenges faced by overfitted personalized models that often struggle to provide good reward signals during RL fine-tuning.
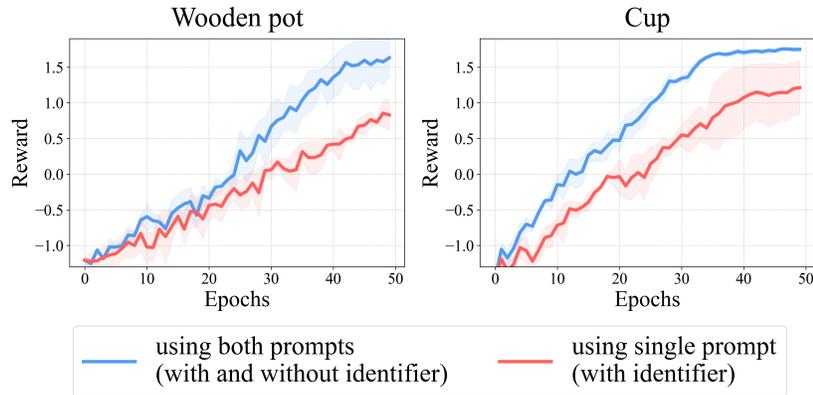

Figure 16. Learning curves of the model trained with a single prompt (i.e., "a [*] [class noun] with pens in it") and the model trained with both prompts (i.e., including "a [class noun] with pens in it"). The solid line and shaded regions represent the mean and standard deviation across three runs, respectively.