Domain-Agnostic Tuning-Encoder for Fast Personalization of Text-To-Image Models

Moab Arar
Tel Aviv University
Tel Aviv, Israel

Rinon Gal
Tel Aviv University, NVIDIA
Tel Aviv, Israel

Yuval Atzmon
NVIDIA
Tel Aviv, Israel

Gal Chechik
Nvidia
Tel Aviv, Israel

Daniel Cohen-Or
Tel Aviv University
Tel Aviv, Israel

Ariel Shamir
Reichman University (IDC)
Herzliya, Israel

Amit H. Bermano
Tel Aviv University
Tel Aviv, Israel

Abstract

Text-to-image (T2I) personalization allows users to guide the creative image generation process by combining their own visual concepts in natural language prompts. Recently, encoder-based techniques have emerged as a new effective approach for T2I personalization, reducing the need for multiple images and long training times. However, most existing encoders are limited to a single-class domain, which hinders their ability to handle diverse concepts. In this work, we propose a domain-agnostic method that does not require any specialized dataset or prior information about the personalized concepts. We introduce a novel contrastive-based regularization technique to maintain high fidelity to the target concept characteristics while keeping the predicted embeddings close to editable regions of the latent space, by pushing the predicted tokens toward their nearest existing CLIP tokens. Our experimental results demonstrate the effectiveness of our approach and show how the learned tokens are more semantic than tokens predicted by unregularized models. This leads to a better representation that achieves state-of-the-art performance while being more flexible than previous methods. [1]

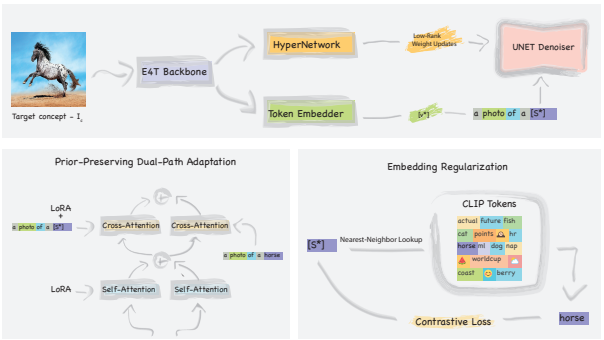

Figure 1: **Method overview.** (top) Our method consists of an encoder with two prediction heads, and a hypernetwork which predicts LoRA-style attention-weight offsets and a word-embedder. (bottom, right) Our embeddings are regularized by using a nearest-neighbour based contrastive loss that pushes them towards real words, but away from the embeddings of other concepts. (bottom, left) We employ a dual-path adaptation approach where each attention branch is repeated twice, once using the soft-embedding and the hypernetwork offsets, and once with the vanilla model and a hard-prompt containing the embedding's nearest neighbor. These branches are linearly blended to better preserve the prior.

1. Introduction

Early personalization methods [3, 15] of text-to-image models rely on the availability of multiple images and require lengthy optimization. An effective alternative is pre-training predictive models for targeting concepts. These approaches train an encoder to predict a text embedding that accurately reconstructs a given desired target concept. Using the obtained embeddings, one can generate scenes portraying the given concept.

Still, such methods face limitations. First, they rely on a single-class domain, which constrains their ability to capture the long-tail distribution of diverse concepts. Second, some approaches necessitate external priors, such as segmentation masks or multi-view input, to effectively capture the characteristics of the target concept while discarding spurious background features.

In this work, we follow E4T [4], an encoder-based approach that requires only (5-15 iteration) of fine-tuning.

[1] Accepted to SIGGRAPH-Asia

Figure 2: Qualitative comparison with existing methods. Our method achieves comparable quality to the state-of-the-art using only a single image and 12 or fewer training steps. Notably, it generalizes to unique objects which recent encoder-based methods struggle with.

In particular, we pre-train an encoder to predict a word-embedding unique to the target concept, and low-rank weight updates (LoRA) [6], quickly adapting the generative model to personalize the target concept. At inference time, we tune the predicted weights and word embedding only, significantly reducing the memory requirement by x4 times and the parameter search space by two orders of magnitude.

To achieve this, we introduce two novel regularizations to avoid over-fitting and preserve the models' capability to generate images from novel prompts. The first uses contrastive learning to ensure the word-embedding lies within the word-embedding manifold. Secondly, we employ a dual-path adaptation approach where each attention branch is repeated twice, once using the soft-embedding and the hyper network offsets and once with the vanilla model, and a hard prompt containing the embedding's nearest neighbor. These branches are linearly blended to preserve the prior better. The latter path keeps the model's prior, while the newly adapted branch adapts to the target concept.

We compare our method to existing encoders and optimization-based approaches and demonstrate that it can achieve high quality and fast personalization across many different domains.

2. Preliminaries

Providing context, we review two recent text-to-image personalization methods: Textual Inversion [3] and E4T [4].

2.1. Textual Inversion

Textual Inversion (TI) [3] personalizes a T2I model by learning a novel word-embedding, $v*$, that will represent a concept visualized in a small (3-5) image set. To find such an embedding, the authors leverage the simple diffusion denoising loss [5]:

$$L_{Diffusion} := \mathbb{E}_{z,y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, y)\|_2^2\right], \quad (1)$$

where ϵ is the unscaled noise sample, ϵ_θ is the denoising network, t is the time step, z_t is an image or latent noised to time t, and c is some conditioning prompt containing an arbitrary string S_* that is mapped to the embedding $v*$.

Once learned, this embedding can be invoked in future prompts (by including the placeholder S_*, *e.g.* "a photo of S_*") in order to generate images of the concept in novel contexts and scenes.

2.2. Encoder for Tuning (E4T):

Optimization-based methods like TI take several minutes to converge. Recently, encoder-based approaches have emerged that train a neural network to directly map an image of a concept to a novel embedding. More concretely, given an input image I_c depicting the concept, the encoder E is trained to predict a suitable embedding: $v_* = E(I; \theta)$. This encoder can be pretrained on a large set of images using the same denoising goal of 1, allowing it to later generalize to new concepts.

In E4T, this encoder is pre-trained on a specific target domain (e.g., human faces, cats, artistic styles). To avoid overfitting and maintain editability, it regularizes the predicted embeddings by constraining them near the embedding of a word describing the domain (e.g., "face," "cat," or "art"). However, this regularization compromises identity preservation, which is later restored through inference-time tuning using a single target image and brief training.

In this work, we extend encoder-based tuning approach to an unrestricted domain.

3. Method

3.1. Architecture Desgin

Our encoder predicts word-embeddings and Low-Rank weight-updates [6] to adapt a pre-trained T2I model for an input image (See top of Fig. 1). We follow the architecture design of E4T architecture with an additional hypernetwork. Further details can be found in appendix A.

3.2. Embedding Regularization

Large Language models handle words by breaking them into a sequence of tokens, subsequently transformed into relevant embeddings $T_{i i=1}^{n}$. During tokenization, each word is linked to one or more high-dimensional vectors, employed as input for transformer-based models.

Our encoder forecasts an embedding, $v* = E(I_c)$, that optimally represents a target concept I_c. Prior work [4] has shown that encoders often employ out-of-distribution embeddings, over-taking attention from other words [17], constraining the ability for subsequent manipulation of personalized concepts via novel prompts. To avoid that, we push the network prediction towards existing word-embedding.

Inspired by [7, 10], we make use of a "nearest-neighbor" contrastive-learning objective with dual goals: (1) push the predicted embedding close to their nearest CLIP tokens, and (2) map different concept images to different embeddings. Concretely, given $v_* = E(I_c)$, we find $\mathbb{N}(v_*)$, the set of nearest CLIP-tokens to v_* in terms of the cosine distance metric. These CLIP tokens, $T_i \in \mathbb{N}(v_*)$ serve as positive examples in the contrastive loss. For every other image $I' \neq I_c$ in the current mini-batch, we use the embedding $v' = E(I')$ as our negative sample. Therefore, our loss is defined by:

$$L_c(v_*) = -\log \frac{\sum_{\mathbb{N}(v_*)} \exp(v_* \cdot T_i/\tau)}{\sum_{\mathbb{N}(v_*)} \exp(v_* \cdot T_i/\tau) + \sum_{v' \neq v_*} \exp(v_* \cdot v'/\tau)} \quad (2)$$

. As opposed to previous methods [7], using the nearest neighbors embeddings as positive samples requires no supervision or prior knowledge on the target domain, canceling the need for a pre-defined list of positive and negative tokens in advance. Finally, we additionally employ an L2-regularization term to prevent the norm of the embedding from increasing significantly.

3.3. Hyper-weights Regularization

The hypernetwork's prediction can also overfit the model to a given image [4]. To address the issue, we propose a modification to the UNET forward pass. We begin by duplicating each block into two copies. The first block uses the original UNET's weights, and for the second, we use the hypernetwork-modulated weights. Moreover, in the first (original weight) branch, we replace our predicted word embeddings with those of the nearest neighbor token. The outputs of the two paths are then linearly blended with a coefficient of α_{blend}. This dual-call approach ensures that one path is free from attention-overfitting, and can thereby strike a balance between capturing the identity and preserving the model's prior knowledge (see Fig 1).

Specifically, given the weight modulations W_Δ and the predicted word embedding v_* from our encoder E, we first identify the nearest hard-token embedding v_h to the model's prediction v_*. We then compose two text prompts, C and C_h, which consist of v_* and v_h respectively. In other words, C and C_h are derived from the same prompt, but one uses the learned embedding while the other uses only real-token ("hard") embeddings.

For each block B of the UNET-denoiser, which receives a feature map $f \in \mathbb{R}^{k \times k \times D}$, text condition C, and weight modulation W_Δ, we modify the block using the dual-path approach:

$$out = \alpha_{blend} \cdot B(f, C, W_\Delta) + (1 - \alpha_{blend}) \cdot B(f, C_h, \emptyset) \quad (3)$$

.

3.4. Inference-time Personalization

As a final step, we follow E4T and employ a brief tuning phase at inference time. While E4T tunes both the model and the encoder at inference time, we find that this proccess requires significant memory (roughly 70GB with the recommended minimal batch size of 16). To reduce this requirement, we note that our model predicts the same embedding and weight decomposition used by LoRA [1, 6]. As such, we can use its output as an initialization for a short LoRA-tuning run, with the addition of an L2-regularization term that aims to keep both weights and embedding close to the original encoder prediction.

4. Experiments

In this section we discuss main results, the experimental details appear in appendix B. In our experiments, we used the CLIP similarity score to evalaute the generated image proximity to a target prompt, or to a given image.

4.1. Comparison with existing methods

We compare our method with existing works using qualitative and quantitative results. Figure 2 showcases the outcomes of multi-domain personalization by comparing different approaches. Specifically, we compare our method with Textual-Inversion [3], Dream-Booth[15], ELITE [18], and popular publicly available LoRA library for Stable Diffusion [1]. For DreamBooth and Textual Inversion, we use the HuggingFace Diffusers implementation [11]. Our results are on-par with full tuning-based methods (DreamBooth, LoRA) and significantly outperform encoder based

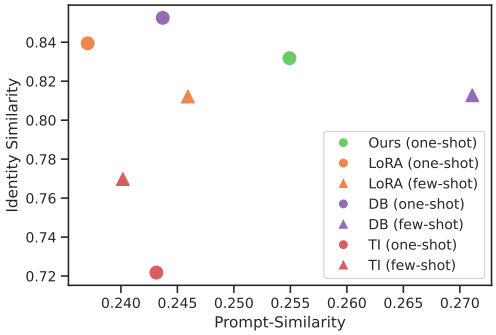

Figure 3: Quantitative comparison to prior work. Our method presents an appealing point on the identity-prompt similarity trade-off curve, while being orders of magnitude quicker than optimization-based methods.

approach of ELITE, even though the latter has access to additional supervision (e.g, segmentation masks). Additional comparison and results generated using our method can be found in the appendix.

Quantitative results are shown in 3. Our method achieves better identity preservation and editability than LoRA, but exhibits a tradeoff when compared to DreamBooth. It outperforms all baselines when they are trained using only a single image. Overall, our approach is competitive with the state-of-the-art while using only a single image and 12 tuning iterations.

5. Ablations

Extensive ablation study appears in Appendix C and D.

References

[1] Low-rank adaptation for fast text-to-image diffusion fine-tuning, 2023.

[2] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 12873–12883. Computer Vision Foundation / IEEE, 2021.

[3] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

[4] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Designing an encoder for fast personalization of text-to-image models. *arXiv preprint arXiv:2302.12228*, 2023.

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[6] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685, 2021.

[7] Ziqi Huang, Tianxing Wu, Yuming Jiang, Kelvin CK Chan, and Ziwei Liu. Reversion: Diffusion-based relation inversion from images. *arXiv preprint arXiv:2303.13495*, 2023.

[8] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. *arXiv preprint arXiv:2212.04488*, 2022.

[9] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020.

[10] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889, 2020.

[11] Suraj Patil and Pedro Cuenca. Huggingface dreambooth implementation. `https://huggingface.co/docs/diffusers/training/dreambooth`, 2022.

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[13] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *arXiv preprint arXiv:2106.05744*, 2021.

[14] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE, 2022.

[15] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. 2022.

[16] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[17] Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. Key-locked rank one editing for text-to-image personalization. *arXiv preprint arXiv:2305.01644*, 2023.

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

A. Architecture Desgin

We adopt the E4T architecture, which features an iterative-refinement design. Specifically, we utilize a pre-trained CLIP [12] ViT-H-14 visual encoder and StableDiffusion's UNET-Encoder as feature-extraction backbones. We extract the spatial features for the given input image from each backbone's last layer. Following E4T, when extracting features from the UNET-Encoder, we provide it with an empty prompt. The features are processed by a convolutional-based network and shared between two prediction heads: a token embedder and a HyperNetwork. The token embedder predicts word embeddings that will be used to represent our target concept I_c. The HyperNetwork predicts weight-modulations for Stable Diffusion's denoising UNET. Next we discuss some important aspects about each prediction head.

HyperNetwork: Capturing fine details of the target concept solely through token embedding is challenging. Earlier methods demonstrated that fine-tuning T2I models achieves high personalization results [15, 3]. Thus, we aim to forecast weight adjustments to refine the denoiser for improved identity preservation. Additionally, we leverage Stable Diffusion [14], boasting nearly a billion parameters. Adapting such a vast number of weights via a HyperNetwork proves computationally unfeasible. Consequently, inspired by prior work [4, 8, 1], we concentrate on predicting low-rank [6, 1] updates for a subset of Stable Diffusion's layers, specifically the attention projection matrices. More specifically, for each concept I_c and each projection matrix W, we predict two matrices , $A \in \mathbb{R}^{D_{in} \times r}$ and $B \in \mathbb{R}^{r \times D_{out}}$, where r is the decomposition rank. The new modulated matrices are:

$$W' = W + \Delta W = W + A \times B \tag{4}$$

.

To avoid breaking the model at the beginning of training, we initialize the prediction layer of the matrix B to zero, and scale ΔW by a constant factor following [6]. We further regularize the weight-offsets by applying $L2$-regularization.

B. Experimental setup

Pre-training: We initiated our experiments by pre-training our model on the ImageNet-1K and Open-Images datasets [16, 9]. ImageNet-1K consists of 1.28 million training images from 1,000 distinct classes. For OpenImages dataset, we crop the largest object from each training image to avoid training on multiple-subjects at the same time. Together, our training data consists of around 3M images. The pre-training phase employed a pre-trained CLIP model with a ViT-H-14 encoder as the backbone architecture. The token-embedder and hyper-network were trained using a learning rate of lr=1e-4 with linear warm-up and cosine-decay scheduling. For ablation purposes, we conducted 50,000 iterations during training. For our final model and comparisons to prior art, we extended the training to 150,000 steps.

Inference-tuning Phase: During the inference-time tuning phase, we used a single-forward pass to obtain the initial prediction of the hyper-weights and word-embedding for the text-to-image model adaptation. Subsequently, we optimized the initial prediction using a learning rate of lr=2e-3 and a balancing factor of $\alpha_{blend} = 0.25$ (see Eq. 3). We found that up to 12 optimization steps were sufficient to achieve satisfactory results for various concepts, compared to the recommended $2,000$ for LoRA-PTI [1, 13].

Evaluation Metric: We follow TI [3] and employ a CLIP text-to-image similarity score as the evaluation metric to assess the proximity of the generated images to the input prompt. To measure identity preservation, we utilized the image-to-image CLIP similarity loss between the single-image training set and the generated results. All reported metrics are based on a pre-trained ViT-B-16 model. Our evaluation set contains 17 images taken from prior work [3, 15, 8]. These cover diverse categories ranging from pets (e.g., dogs) to personal items (e.g., backpacks) and even buildings.

C. The importance of contrastive regularization

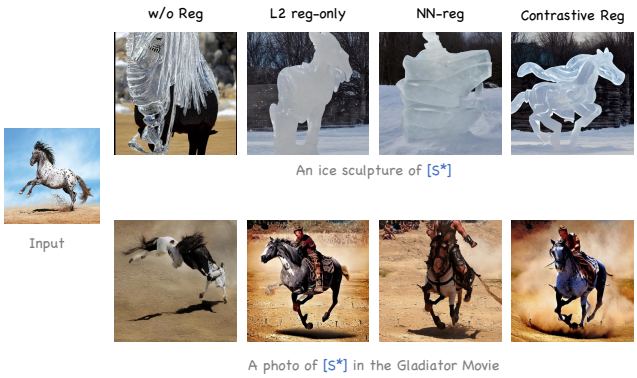

Figure 4: The effects of removing or changing the embedding regularization. Removal of regularization leads to overfitting or mode collapse with poor quality results. Naive regularizations tend to struggle with preserving the concept details. Our contrastive-based regularization can achieve a tradeoff between the two.

Our approach utilizes contrastive learning to improve the quality of predicted embeddings. To visualize the bene-

fit of this regularization, we train our model in four settings: First, without any regularization. Second, we omit all regularization except for the L2 loss on the predicted embedding. Third, we replace the contrastive loss with one that minimizes the cosine-distance between predicted embeddings and their nearest neighbor - a loss inspired by the codebook losses employed in VQGAN [2]. Finally, we use our proposed contrastive-based alternative.

As seen in Fig 4, incorporating our contrastive-based loss improves results. In particular, omitting any regularization tends to overfit the input image. For example, in the generated image of "A photo of [S*] in the gladiator movie," the word gladiator is overlooked. And the model overfits the predicted token. On the other hand, using our contrastive loss, the generated photo faithfully describes the input prompt while preserving features of the target concept (i.e., the horse). The contrastive loss function also helps to prevent mode collapse by repelling tokens of different images via negative samples. For example, unlike the contrastive-based method, the nearest-neighbor approach does not address mode collapse. It yields less favorable results (See Fig 4).

D. Ablation Analysis

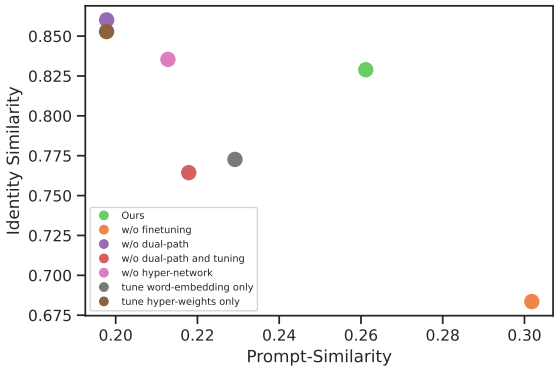

Figure 5: Ablation study results. Removing regularization typically leads to quick overfitting, where editability suffers. Skipping the fine-tuning step harms identity preservation, in line with E4T [4].

We conduct an ablation study to better understand the importance of each component in our method. The results appear in Fig. 5. We examine the following setups: removing the dual-path regularization approach, skipping the fine-tuning step, and omitting the hypernetwork branch. We observe that the final tuning step is crucial, inline with the observation from E4T. In particular, when using our baseline without finetuning, we witness a 20% drop in the object similarity metric. Turning off the dual-path during tuning harms prompt-to-image alignment by nearly 30%, suggesting heavy overfitting. Hence, we can conclude that the dual-path approach can successfully preserve the prior and diminish overfitting.

Another important component of our method is the hyper-network, which predicts weight modulations to calibrate the generator with our target concept. In our ablation study, we found that omitting the hyper-network at training time negatively impacts the alignment of the generated images with the text prompts. We believe this is because the network must encode more information about the object in the word-embedding, causing attention-overfitting as described in the method sections.

E. Additional qualitative results

We include additional results of our method in Fig 6 and Fig 7.

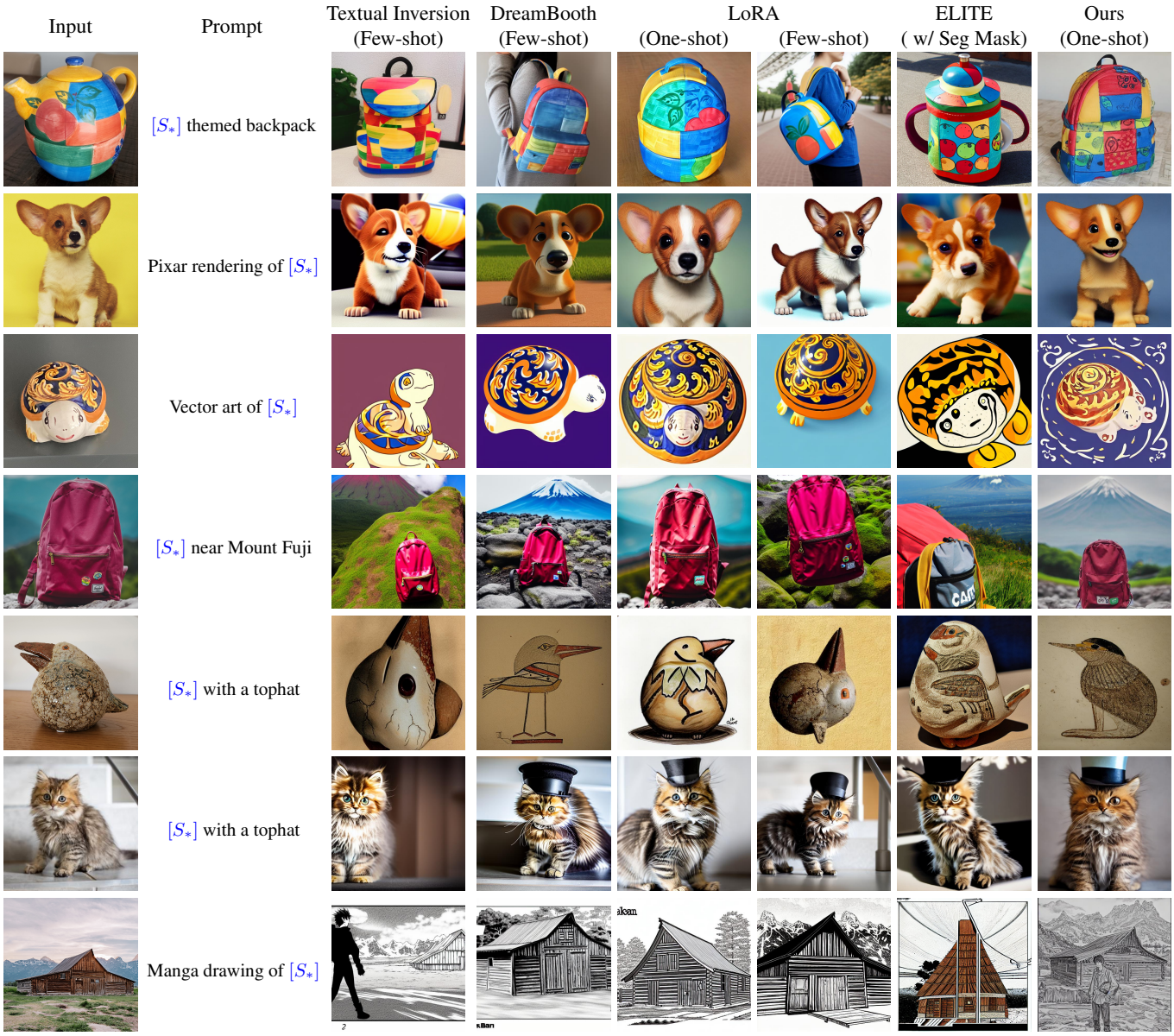

Figure 6: Qualitative comparison with existing methods. Our method achieves comparable quality to the state-of-the-art using only a single image and 12 or fewer training steps. Notably, it generalizes to unique objects which recent encoder-based methods struggle with.

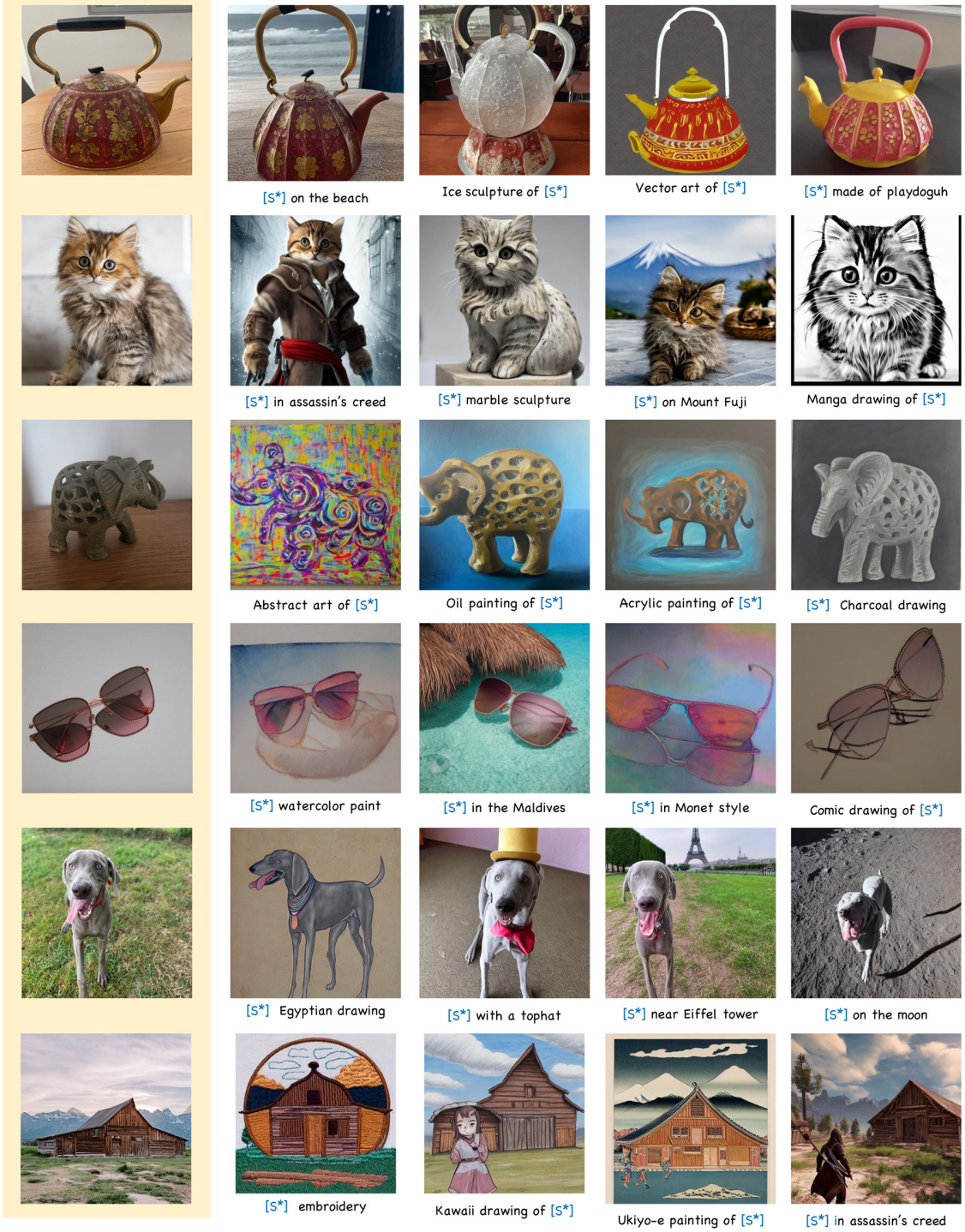

Figure 7: Additional qualitative results generated using our method. The left-most column shows the input image, followed by 4 personalized generations for each subject.