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

MOAB ARAR, Tel Aviv University, Israel
RINON GAL, Tel Aviv University, NVIDIA, Israel
YUVAL ATZMON, NVIDIA, Israel
GAL CHECHIK, NVIDIA, Israel
DANIEL COHEN-OR, Tel Aviv University, Israel
ARIEL SHAMIR, Reichman University (IDC), Israel
AMIT H. BERMANO, Tel Aviv University, Israel

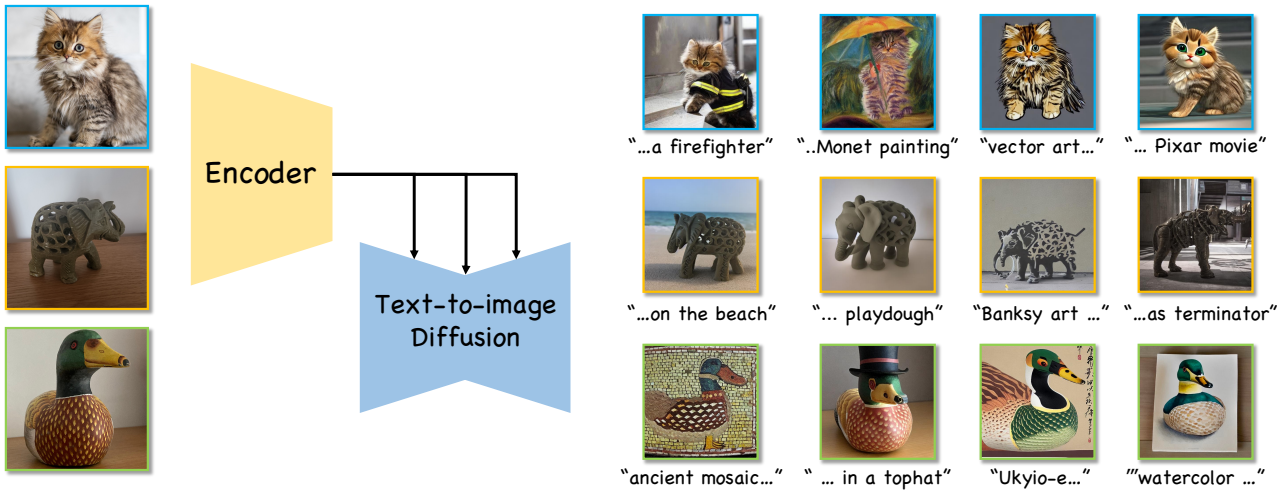

Fig. 1. Our domain-agnostic tuning-encoder can personalize a text-to-image diffusion model to a given concept using 12 or fewer training steps, allowing for general one-shot inference-time tuning. The personalized models are used to generate images of the concept in new settings using natural language prompts.

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 INTRODUCTION

The rapid advancement of generative models has revolutionized content creation, enabling effortless generation of diverse artworks. Part of their true potential lies in personalization, allowing users to tailor outputs to unique personal concepts. Personalizing a model involves customizing it to capture and manifest unique characteristics of personal belongings, memories, or self-portraits. However, early personalization methods rely on the availability of multiple images or require lengthy optimization steps.

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 [Gal et al. 2023], an approach which leverages the encoder as a form of initialization for brief (5-15 iteration) fine-tuning. E4T trains an encoder for each individual domain, and requires roughly 70GB of VRAM for inference-time tuning. Our approach can tackle multiple domains, and reduces inference-time memory requirements. We consider two goals while designing our encoder: (1) the ability to edit the target concepts, and (2) the ability

to faithfully capture distinguishable characteristics of the target. We achieve the first goal by regularizing the model to predict words within the editable region of the generative model. Unlike prior single-domain methods, we do not rely on a coarse description of the target domain. Instead, we use a contrastive-based approach to push the predicted embedding toward meaningful regions in the word embedding space. Intuitively, ensuring the prediction is near words that semantically describe the concept class will better preserve the model's prior knowledge of the concept class.

For the second goal, we introduce a hyper-network to capture the distinctive features of the target concepts with higher fidelity. To ensure a manageable model size, we employ a strategy of predicting a low-rank decomposition of the weights of the UNET-denoiser model, following the approach outlined in Hu et al. [2021] and Gal et al. [2023]. Finally, the joint embedding and hyper-network predictions are used to initialize a regularized LoRA training process, requiring 12 or fewer optimization steps. Importantly, this reduces memory requirements from roughly 70GB to fewer than 30GB and shortens training and inference times.

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 RELATED WORK

Text-driven image generation using diffusion models. Text-to-image synthesis has made significant progress in recent years, driven mainly by pre-trained diffusion models [Ho et al. 2020a] and especially by large models [Balaji et al. 2022; Nichol et al. 2021; Ramesh et al. 2022; Rombach et al. 2022] trained on web-scale data like [Schuhmann et al. 2021]. Our approach builds upon these pre-trained models to extend their vocabulary and generate personalized concepts. Specifically, we use the Stable-Diffusion model [Rombach et al. 2022]. We expect it to generalize to diffusion-based generators with similar attention-based architectures [Saharia et al. 2022].

Text-based image editing. Following the success of CLIP-based editing methods [Bar-Tal et al. 2022; Gal et al. 2021; Michel et al. 2021; Patashnik et al. 2021], a large body of work sought to leverage the power of recent large-scale text-to-image models [Balaji et al. 2022; Kang et al. 2023; Ramesh et al. 2022; Rombach et al. 2022; Saharia et al. 2022; Sauer et al. 2023] in order to manipulate images using text-based guidance. Prompt-to-Prompt [Hertz et al. 2022] propose a method for manipulating *generated* images by re-using an initial prompt's attention masks. In a follow-up work, Mokady et al. [2022] extend this approach to real images by encoding them into the null-conditioning space of classifier-free guidance [Ho and Salimans 2021]. Tumanyan et al. [2022] and Parmar et al. [2023] extract reference attention maps or features using DDIM [Song et al. 2020] based reconstructions. These are then used to preserve image structure under new prompts. Others train an instruction-guided image-to-image translation network using synthetic data [Brooks et al. 2023] or tune the model to reconstruct an image and use conditioning-space walks to modify it [Kawar et al. 2022].

Common to these image-editing approaches is a desire to preserve the content of the original image. In contrast, our method deals with model personalization which aims to capture a concept for later use in novel scenes. There, the aim is to learn the semantics and apperance of a subject, but not its specific structure in the image.

Inversion. In the context of Generative Adversarial Networks (GANs, [Goodfellow et al. 2014]), inversion is the task of finding a latent representation that will reproduce a specific image when passed through a pre-trained generator [Xia et al. 2021; Zhu et al. 2016]. There, methods are split into two main camps. In the first are optimization methods, which iterative search the latent space for a code that synthesizes an image with some minimal reconstruction loss [Abdal et al. 2019, 2020; Gu et al. 2020; Zhu et al. 2020a]. In the second are encoder based methods, which train a neural network to directly predict such latents [Bai et al. 2022; Parmar et al. 2022; Pidhorskyi et al. 2020; Richardson et al. 2020; Tov et al. 2021; Wang et al. 2022; Zhu et al. 2020b].

With diffusion models, the inversion latent space can be the initial noise map that will later be denoised into a given target [Dhariwal and Nichol 2021; Ramesh et al. 2022; Song et al. 2020]. In a more recent line of work, inversion has been used to refer to finding a conditioning code that can be used to synthesize novel images of a given concept [Gal et al. 2022]. There, the goal is not to recreate a specific image, but to capture the semantics of a concept outlined in one or more target images and later re-create it in new scenes. Our approach similarly aims to encode a concept.

Personalization. Personalization methods aim to tune a model to a specific individual target. Often, the goal is to combine some large-scale prior knowledge with unique information associated with an end-user. These can include personalized recommendation systems [Amat et al. 2018; Benhamdi et al. 2017; Cho et al. 2002; Martinez et al. 2009], federated learning [Fallah et al. 2020; Jiang et al. 2019; Mansour et al. 2020; Shamsian et al. 2021], or the creation of generative models tuned on specific scenes or individuals [Alaluf et al. 2021; Bau et al. 2019; Cao et al. 2022; Cohen et al. 2022; Dinh et al. 2022; Nitzan et al. 2022; Roich et al. 2021]. In text-to-image personalization, the goal is to teach pre-trained models to synthesize novel images of a specific target concept, guided by natural language prompts. Initial work in this field employed direct optimization approaches, either tuning a set of text embeddings to describe the concept [Gal et al. 2022], modifying the denoising network itself [Ruiz et al. 2022], or a mixture of both [sim 2023; Kumari et al. 2022; Tewel et al. 2023]. However, such optimization-based approaches require lengthy training sessions, typically requiring dozens of minutes for every concept.

More recently, encoder-based approaches emerged [Gal et al. 2023; Shi et al. 2023; Wei et al. 2023], which train a neural network to predict some latent representation that can be injected into the network to synthesize new images of the concept. These either require subject-specific segmentation masks [Wei et al. 2023] or use single-domain training to both regularize the model and allow it to infer the target from the single image [Gal et al. 2023; Shi et al. 2023]. In an alternative approach, a model can be trained to synthesize novel images from dual conditions: a text prompt, and a set of images depicting the target [Chen et al. 2023]. However, this approach is based on apprenticehsip learning, where the model is trained on outputs from half a million pre-trained personalized models. Such

an approach therefore requires roughly 14 A100 GPU-years, making it infeasible for most practitioners.

Our method follows the encoder-based approach, but extends it beyond the single-domain without use of any segmentation masks or additional labels. Moreover, compared to prior encoder-based tuning approaches [Gal et al. 2023], our tuning-phase is quicker and has reduced memory overhead.

3 PRELIMINARIES

To put our contribution in context, we begin with an overview of two recent text-to-image personalization approaches: Textual Inversion [Gal et al. 2022] and E4T [Gal et al. 2023] which serve as a basis for our work.

3.1 Textual Inversion

Textual Inversion (TI) introduced the topic of text-to-image (T2I) personalization, where a pre-trained T2I diffusion model is taught how to reason about unique, user-provided concepts which were unseen during training. In TI, the authors propose to tackle this task 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 [Ho et al. 2020b]:

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

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.

3.2 Encoder for Tuning (E4T):

Although optimization-based approaches like TI can reconstruct the target concept, they require many iterations to converge. Indeed, personalizing a model with TI typically requires dozens of minutes even on commercial-grade GPUs. 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 eq. (1), allowing it to later generalize to new concepts.

In E4T, this encoder is pre-trained on a single target domain (*e.g.* human faces, cats or artistic styles). However, in order to prevent overfitting and preserve editability, it regularizes the predicted embeddings by restricting them to a region close to the embedding of a word describing the single domain (*e.g.* "face", "cat" or "art"). This regularization comes at the cost of identity preservation, which the authors later restore through an inference-time tuning session using a single target image of the concept and a few seconds of training.

Our goal is to extend this encoder-based tuning approach to an unrestricted domain, allowing a user to quickly personalize a model even for rare concepts for which large training sets may not exist.

4 METHOD

4.1 Architecture Desgin

We adopt the E4T architecture, which features an iterative-refinement design. Specifically, we utilize a pre-trained CLIP 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: It is challenging to capture the fine details of the target concept by using only a token embedding. Previous works showed that modulating subsets of the denoiser weights can improve reconstruction quality with minor harm to the model's prior. Therefore, we seek to predict a set of weight modulations to help tune the denoiser for better identity preservation. Moreover, we make use of Stable Diffusion [Rombach et al. 2022], which consists of roughly a billion parameters. Adapting so many weights using a HyperNetwork is computationally infeasible. Hence, we follow prior art [sim 2023; Gal et al. 2023; Kumari et al. 2022] and focus on predicting modulations for a subset of Stable Diffusion's layers, and specifically for the attention projection matrices. However, Stable Diffusion contains 96 such matrices, each containing an average of 715, 946 parameters. Predicting such large matrices is still challenging. Instead, we predict decomposed weights of the same form as Low-Rank Adaptation (LoRA) [Hu et al. 2021], where each weight, $W \in \mathbb{R}^{D_{in}\times D_{out}}$, is modulated by injecting trainable rank decomposition 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 the new modulated matrices are:

$$W' = W + \Delta W = W + A \times B \tag{2}$$

.

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 [Hu et al. 2021]. We further regularize the weight-offsets by applying $L2$-regularization.

4.2 Embedding Regularization

Large Language models are trained on a finite dictionary composed of tokens. Particularly, these models process words by dividing them into a sequence of tokens from the dictionary, which are then converted into appropriate embeddings $\{T_i\}_{i=1}^n$. In this tokenization process, each word is mapped to one or more high-dimensional vectors, which is used as input for transformer-based model.

Our encoder's objective is to predict an embedding, $v* = E(I_c)$, that best describes a target-concept I_c. Previous works [Gal et al. 2023] have shown that in under-constrained settings, encoders tend to use out-of-distribution embeddings. These tend to draw attention

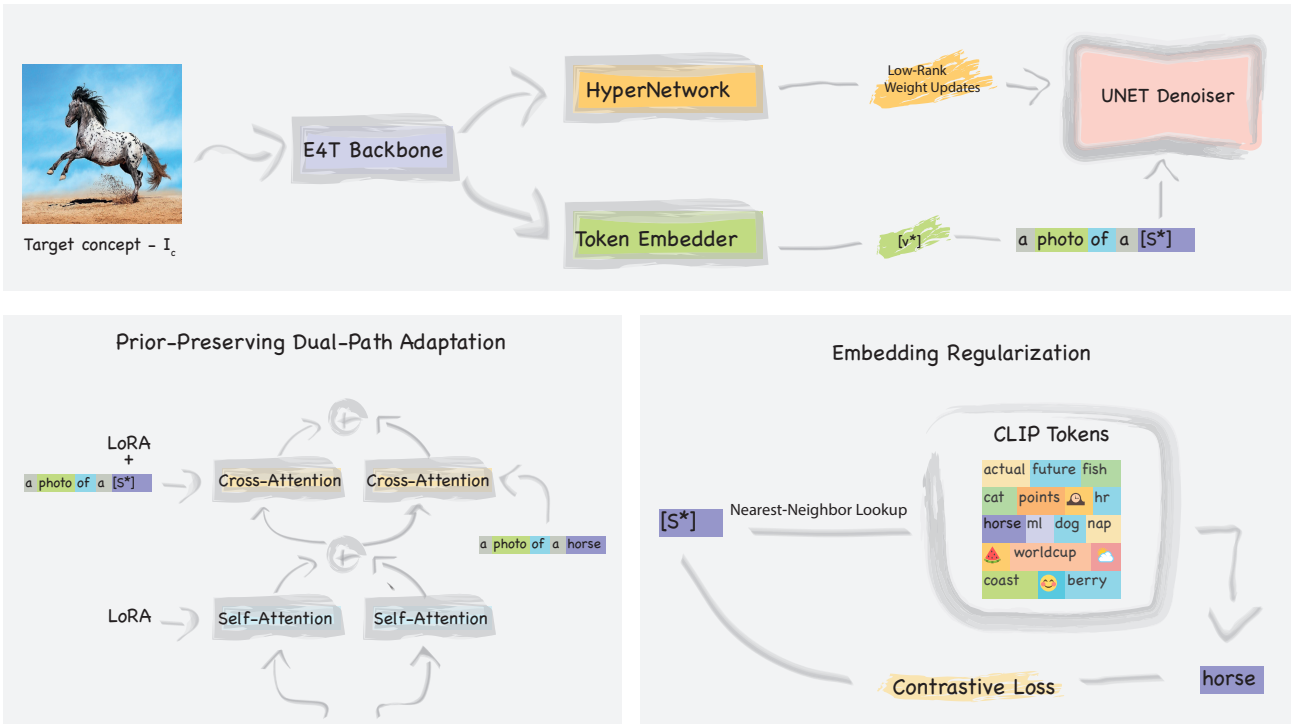

Fig. 2. **Method overview.** (top) Our method consists of a feature-extraction backbone which follows the E4T approach and uses a mix of CLIP-features from the concept image, and denoiser-based features from the current noisy generation. These features are fed into an embedding prediction head, and a hypernetwork which predicts LoRA-style attention-weight offsets. (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.

away from other words [Tewel et al. 2023], limiting the ability to later manipulate the personalized concept via novel prompts. To prevent this attention-overfitting, we could use existing token embeddings to describe I_c. While these tokens are within the training distribution and hence editable, they are not expressive enough to capture personal concepts. We thus relax this hard constraint and predict embeddings *close* to existing tokens. Intuitively, constraining $E(I_c)$ near semantically related words balances the trade-off between reconstruction and editing. However, unlike in single-domain encoders, where a coarse description of the domain exists and is known a-priori, in our setting there could be many semantically different words describing different concepts in the training data. Moreover, the domain encountered during inference may differ from those observed in training.

Inspired by [Huang et al. 2023; Miech et al. 2020], 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:

$$L_c(v_*) = -\log \frac{\sum_{\mathbb{N}(v_*)} \exp(v_* \cdot T_i/\tau)}{\sum_{\mathbb{N}(v_*)} \exp(v_* \cdot T_i/\tau) + \sum_{v' \neq v_*} \exp(v_* \cdot v'/\tau)} \quad (3)$$

. As opposed to previous methods [Huang et al. 2023], 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 embeddings from increasing significantly:

$$L_{L2}(v_*) = ||v_*||^2 \quad (4)$$

4.3 Hyper-weights Regularization

Our encoder also contains a hypernetwork branch, whose predicted weights can similarly overfit the model to a given image [Gal et al. 2023]. To address the issue of overfitting caused by the hypernetwork predictions, 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

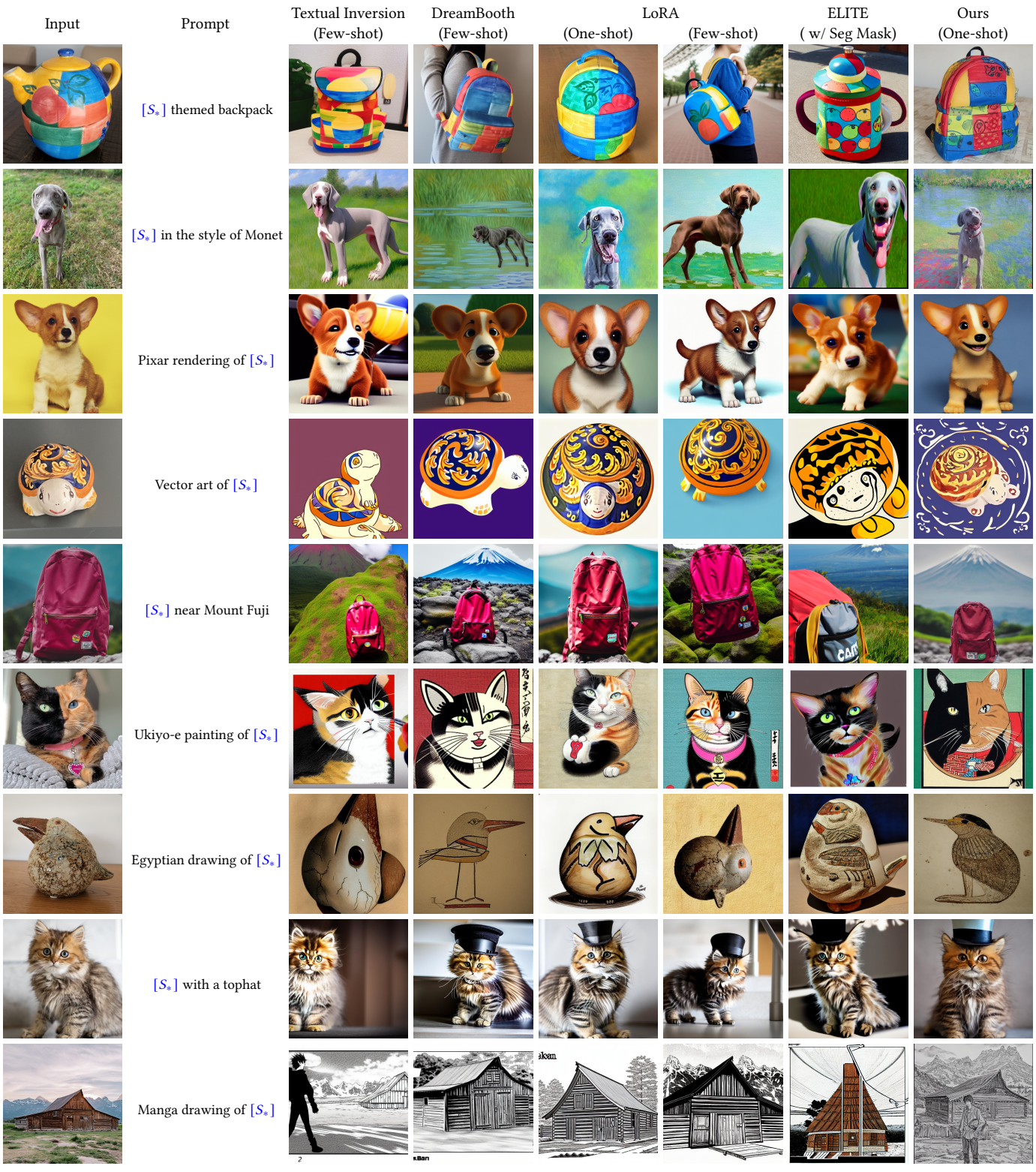

Fig. 3. Qualitative comparison with existing methods. Our method achieves comparable quality to the state-of-the-art using only a single image and 12 or fewer training steps. Notably, it generalizes to unique objects which recent encoder-based methods struggle with.

(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 2).

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:

$$out = \alpha_{blend} \cdot B(f, C, W_\Delta) + (1 - \alpha_{blend}) \cdot B(f, C_h, \emptyset) \quad (5)$$

.

4.4 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 [sim 2023; Hu et al. 2021]. 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.

5 EXPERIMENTS

5.1 Experimental setup

Pre-training: We initiated our experiments by pre-training our model on the ImageNet-1K dataset [Russakovsky et al. 2015], which consists of 1.28 million training images from 1,000 distinct classes. 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 different learning rates: lr=1e-4 and lr=1e-5, respectively. 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. 5). 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 [sim 2023; Roich et al. 2021].

Evaluation Metric: We follow TI [Gal et al. 2022] 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 [Gal et al. 2022; Kumari et al. 2022; Ruiz et al. 2022]. These cover diverse categories ranging from pets (e.g., dogs) to personal items (e.g., backpacks) and even buildings.

5.2 The importance of contrastive regularization

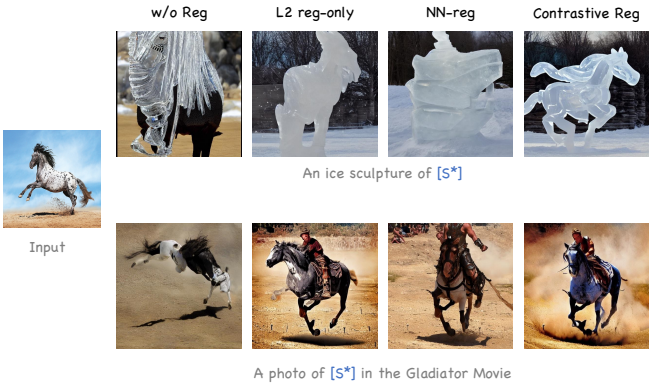

Fig. 4. The effects of removing or changing the embedding regularization. Removal of regularization leads to overfitting or mode collapse with poor quality results. Naïve regularizations tend to struggle with preserving the concept details. Our contrastive-based regularization can achieve a tradeoff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

Our approach utilizes contrastive learning to improve the quality of predicted embeddings. To visualize the benefit 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 [Esser et al. 2021]. 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).

5.3 Comparison with existing methods

We commence our evaluation with a qualitative analysis, demonstrating the ability of our method to capture a remarkable level

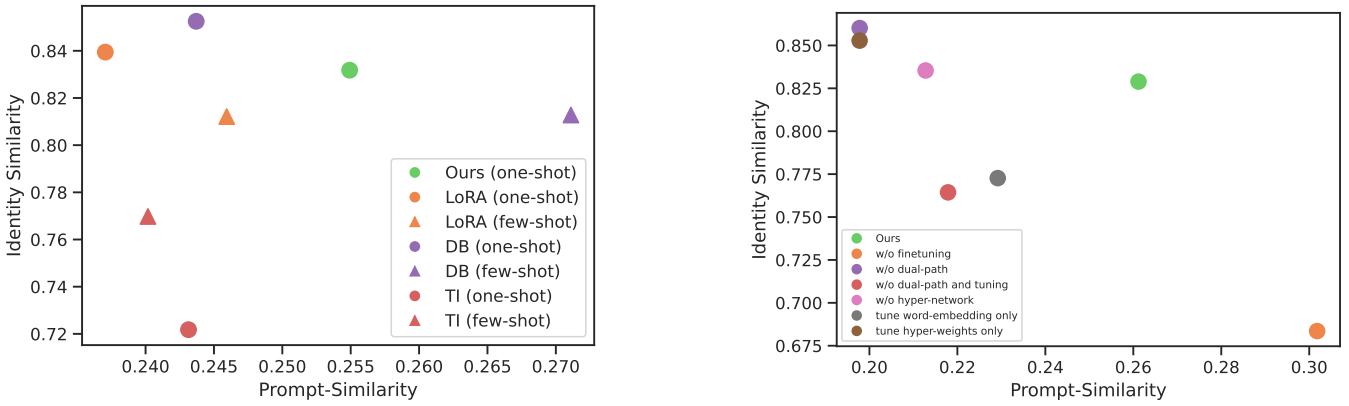

Fig. 5. Quantitative evaluation results. (a) Comparisons to prior work. Our method presents an appealing point on the identity-prompt similarity trade-off curve, while being orders of magnitude quicker than optimization-based methods. (b) Ablation study results. Removing regularization typically leads to quick overfitting, where editability suffers. Skipping the fine-tuning step harms identity preservation, in line with E4T [Gal et al. 2023].

of detail using a single image and a fraction of the training steps. Figure 3 showcases the outcomes of multi-domain personalization by comparing different approaches. Specifically, we compare our method with Textual-Inversion [Gal et al. 2022], Dream-Booth[Ruiz et al. 2022], and popular publicly available LoRA library for Stable Diffusion [sim 2023]. We also compare our method to ELITE [Wei et al. 2023], a state-of-the-art multi-domain personalization encoder. For DreamBooth and Textual Inversion, we use the HuggingFace Diffusers implementation [Patil and Cuenca 2022]. Our results are on-par with full tuning-based methods (DreamBooth, LoRA) and significantly outperform the purely-encoder based approach of ELITE, even though the latter has access to additional supervision in the form of segmentation masks. Notably, all tuning-based methods require access to multiple images of the target concept, while our approach utilizes only a single input. Additional results generated using our method can be found in fig. 6.

Next, we compare our method to the tuning-based approaches using the CLIP-based metrics. Results are shown in fig. 5a. Our method achieves better identity preservation and editability than LoRA, but exhibits a tradeoff when compared to DreamBooth. Notably, 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.

5.4 Ablation Analysis

We conduct an ablation study to better understand the importance of each component in our method. 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.

6 LIMITATIONS

While our approach can extend existing tuning-encoders to multi-class domains, it is still limited by our training data. As such, domains which are poorly represented in the dataset may be hard to encode. As such, a model trained on ImageNet may struggle with cluttered scenes or with human faces.

We believe this limitation can be overcome by training on more general, large-scale datasets such as LAION [Schuhmann et al. 2021]. However, such an investigation is beyond our resources.

While our method can work across a more general domain, it still requires a tuning-step to increase downstream similarity. However, as the memory requirements and iterations required for such tuning-approaches decreases, they become negligible compared to the time required for synthesis.

7 CONCLUSION

We presented a method for generalizing the tuning-encoder approach beyond a single class domain. Our approach restricts overfitting by ensuring predicted embeddings lie close to the real word domain, and by utilizing a dual-pass approach where the network

blends predictions from hard- and soft-prompts. This in turn allows us to quickly personalize a model at inference-time, speeding up personalization by two orders of magnitude compared to optimization-based approaches.

In the future, we hope to further reduce the tuning requirements so that our method can be used on consumer-grade GPUs, allowing end-users to quickly personalize models on their own machine.

ACKNOWLEDGMENTS

This work was partially supported by Len Blavatnik and the Blavatnik family foundation, the Deutsch Foundation, the Yandex Initiative in Machine Learning, BSF (grant 2020280) and ISF (grants 2492/20 and 3441/21). The first author is supported by the Miriam and Aaron Gutwirth scholarship.

REFERENCES

2023. Low-rank Adaptation for Fast Text-to-Image Diffusion Fine-tuning.

Rameen Abdal, Yipeng Qin, and Peter Wonka. 2019. Image2stylegan: How to embed images into the stylegan latent space?. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4432–4441.

Rameen Abdal, Yipeng Qin, and Peter Wonka. 2020. Image2stylegan++: How to edit the embedded images?. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8296–8305.

Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit H. Bermano. 2021. HyperStyle: StyleGAN Inversion with HyperNetworks for Real Image Editing. arXiv:2111.15666 [cs.CV]

Fernando Amat, Ashok Chandrashekar, Tony Jebara, and Justin Basilico. 2018. Artwork Personalization at Netflix. In *Proceedings of the 12th ACM Conference on Recommender Systems* (Vancouver, British Columbia, Canada) *(RecSys '18)*. Association for Computing Machinery, New York, NY, USA, 487–488. https://doi.org/10.1145/3240323.3241729

Qingyan Bai, Yinghao Xu, Jiapeng Zhu, Weihao Xia, Yujiu Yang, and Yujun Shen. 2022. High-fidelity GAN inversion with padding space. In *European Conference on Computer Vision*. Springer, 36–53.

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. 2022. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324* (2022).

Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. 2022. Text2LIVE: Text-Driven Layered Image and Video Editing. *arXiv preprint arXiv:2204.02491* (2022).

David Bau, Hendrik Strobelt, William Peebles, Jonas Wulff, Bolei Zhou, Jun-Yan Zhu, and Antonio Torralba. 2019. Semantic Photo Manipulation with a Generative Image Prior. 38, 4 (2019). https://doi.org/10.1145/3306346.3323023

Soulef Benhamdi, Abdesselam Babouri, and Raja Chiky. 2017. Personalized recommender system for e-Learning environment. *Education and Information Technologies* 22, 4 (2017), 1455–1477.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. 2023. InstructPix2Pix: Learning to Follow Image Editing Instructions. In *CVPR*.

Chen Cao, Tomas Simon, Jin Kyu Kim, Gabe Schwartz, Michael Zollhoefer, Shunsuke Saito, Stephen Lombardi, Shih-en Wei, Danielle Belko, Shoou-i Yu, Yaser Sheikh, and Jason Saragih. 2022. Authentic Volumetric Avatars From a Phone Scan. *ACM Trans. Graph.* (2022).

Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Rui, Xuhui Jia, Ming-Wei Chang, and William W. Cohen. 2023. Subject-driven Text-to-Image Generation via Apprenticeship Learning. *ArXiv* abs/2304.00186 (2023).

Yoon Ho Cho, Jae Kyeong Kim, and Soung Hie Kim. 2002. A personalized recommender system based on web usage mining and decision tree induction. *Expert systems with Applications* 23, 3 (2002), 329–342.

Niv Cohen, Rinon Gal, Eli A. Meirom, Gal Chechik, and Yuval Atzmon. 2022. "This is my unicorn, Fluffy": Personalizing frozen vision-language representations. In *European Conference on Computer Vision (ECCV)*.

Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems* 34 (2021), 8780–8794.

Tan M Dinh, Anh Tuan Tran, Rang Nguyen, and Binh-Son Hua. 2022. Hyperinverter: Improving stylegan inversion via hypernetwork. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11389–11398.

Patrick Esser, Robin Rombach, and Björn Ommer. 2021. Taming Transformers for High-Resolution Image Synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 12873–12883. https://doi.org/10.1109/CVPR46437.2021.01268

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. 2020. Personalized federated learning: A meta-learning approach. *arXiv preprint arXiv:2002.07948* (2020).

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

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

Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. 2021. Stylegan-nada: Clip-guided domain adaptation of image generators. *arXiv preprint arXiv:2108.00946* (2021).

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).

Jinjin Gu, Yujun Shen, and Bolei Zhou. 2020. Image Processing Using Multi-Code GAN Prior. arXiv:1912.07116 [cs.CV]

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-prompt image editing with cross attention control. (2022).

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020a. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* 33 (2020), 6840–6851.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020b. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* 33 (2020), 6840–6851.

Jonathan Ho and Tim Salimans. 2021. Classifier-Free Diffusion Guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *ArXiv* abs/2106.09685 (2021).

Ziqi Huang, Tianxing Wu, Yuming Jiang, Kelvin CK Chan, and Ziwei Liu. 2023. ReVersion: Diffusion-Based Relation Inversion from Images. *arXiv preprint arXiv:2303.13495* (2023).

Yihan Jiang, Jakub Konečnỳ, Keith Rush, and Sreeram Kannan. 2019. Improving federated learning personalization via model agnostic meta learning. *arXiv preprint arXiv:1909.12488* (2019).

Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. 2023. Scaling up GANs for Text-to-Image Synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2022. Imagic: Text-Based Real Image Editing with Diffusion Models. *arXiv preprint arXiv:2210.09276* (2022).

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2022. Multi-Concept Customization of Text-to-Image Diffusion. *arXiv preprint arXiv:2212.04488* (2022).

Yishay Mansour, Mehryar Mohri, Jae Ro, and Ananda Theertha Suresh. 2020. Three approaches for personalization with applications to federated learning. *arXiv preprint arXiv:2002.10619* (2020).

Ana Belen Barragans Martinez, Jose J Pazos Arias, Ana Fernandez Vilas, Jorge Garcia Duque, and Martin Lopez Nores. 2009. What's on TV tonight? An efficient and effective personalized recommender system of TV programs. *IEEE Transactions on Consumer Electronics* 55, 1 (2009), 286–294.

Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. 2021. Text2Mesh: Text-Driven Neural Stylization for Meshes. *arXiv preprint arXiv:2112.03221* (2021).

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

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Null-text Inversion for Editing Real Images using Guided Diffusion Models. *arXiv preprint arXiv:2211.09794* (2022).

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741* (2021).

Yotam Nitzan, Kfir Aberman, Qiurui He, Orly Liba, Michal Yarom, Yossi Gandelsman, Inbar Mosseri, Yael Pritch, and Daniel Cohen-Or. 2022. MyStyle: A Personalized Generative Prior. *arXiv preprint arXiv:2203.17272* (2022).

Gaurav Parmar, Yijun Li, Jingwan Lu, Richard Zhang, Jun-Yan Zhu, and Krishna Kumar Singh. 2022. Spatially-Adaptive Multilayer Selection for GAN Inversion and Editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11399–11409.

Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. 2023. Zero-shot Image-to-Image Translation. arXiv:2302.03027 [cs.CV]

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. *arXiv preprint arXiv:2103.17249* (2021).

Suraj Patil and Pedro Cuenca. 2022. HuggingFace DreamBooth Implementation. https://huggingface.co/docs/diffusers/training/dreambooth.

Stanislav Pidhorskyi, Donald A Adjeroh, and Gianfranco Doretto. 2020. Adversarial Latent Autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14104–14113.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* (2022).

Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. 2020. Encoding in Style: a StyleGAN Encoder for Image-to-Image Translation. *arXiv preprint arXiv:2008.00951* (2020).

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

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. 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*. IEEE, 10674–10685. https://doi.org/10.1109/CVPR52688.2022.01042

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2022. DreamBooth: Fine Tuning Text-to-image Diffusion Models for Subject-Driven Generation. (2022).

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. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115, 3 (2015), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *arXiv preprint arXiv:2205.11487* (2022).

Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. 2023. StyleGAN-T: Unlocking the Power of GANs for Fast Large-Scale Text-to-Image Synthesis. *International Conference on Machine Learning* abs/2301.09515. https://arxiv.org/abs/2301.09515

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114* (2021).

Aviv Shamsian, Aviv Navon, Ethan Fetaya, and Gal Chechik. 2021. Personalized federated learning using hypernetworks. In *International Conference on Machine Learning*. PMLR, 9489–9502.

Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. 2023. InstantBooth: Personalized Text-to-Image Generation without Test-Time Finetuning. arXiv:2304.03411 [cs.CV]

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. 2023. Key-Locked Rank One Editing for Text-to-Image Personalization. *arXiv preprint arXiv:2305.01644* (2023).

Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. 2021. Designing an Encoder for StyleGAN Image Manipulation. *arXiv preprint arXiv:2102.02766* (2021).

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. 2022. Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation. *arXiv preprint arXiv:2211.12572* (2022).

Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. 2022. High-Fidelity GAN Inversion for Image Attribute Editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. 2021. GAN Inversion: A Survey. arXiv:2101.05278 [cs.CV]

Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. 2020b. In-domain gan inversion for real image editing. *arXiv preprint arXiv:2004.00049* (2020).

Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. 2016. Generative visual manipulation on the natural image manifold. In *European conference on computer vision*. Springer, 597–613.

Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. 2020a. Improved StyleGAN Embedding: Where are the Good Latents? arXiv:2012.09036 [cs.CV]

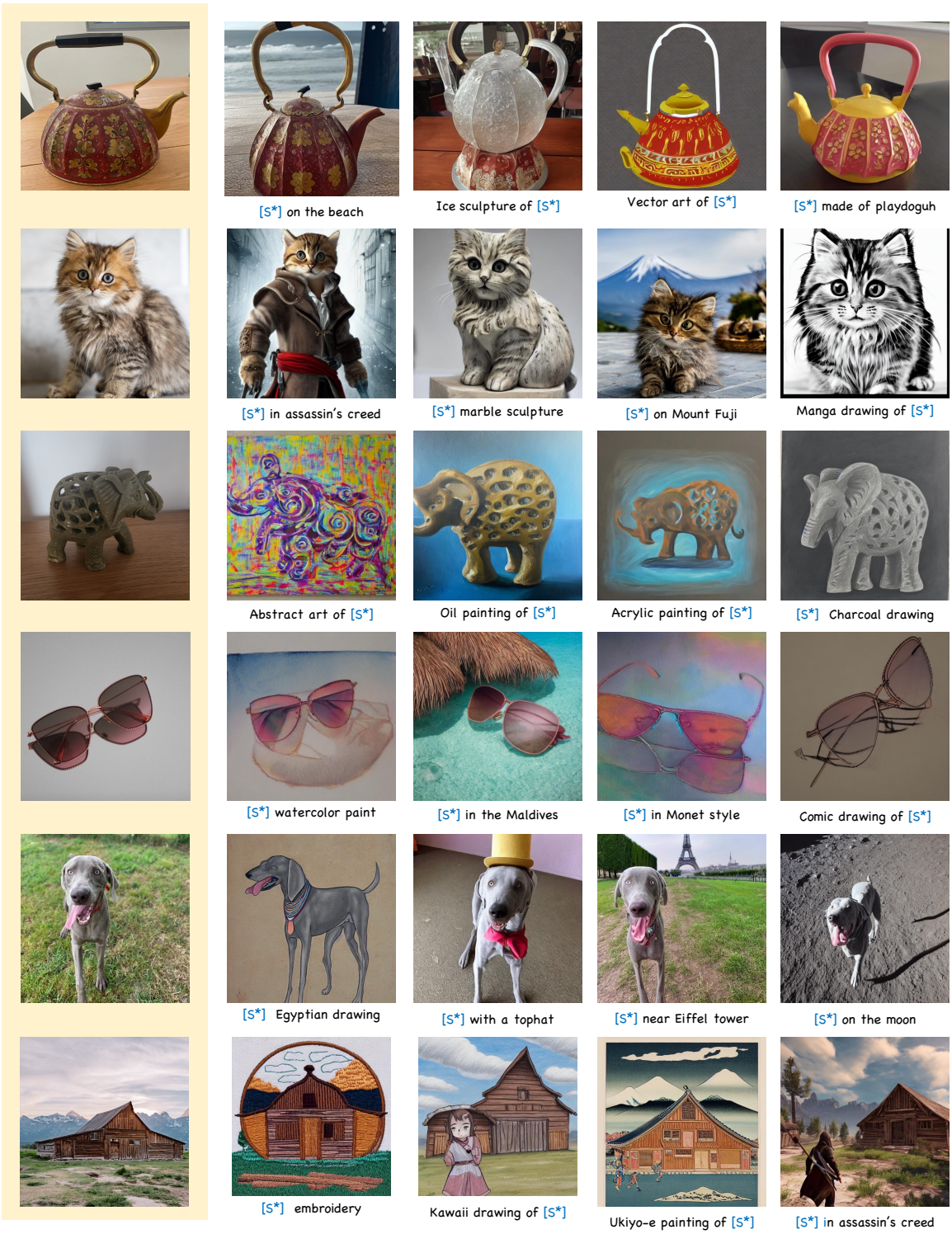

Fig. 6. Additional qualitative results generated using our method. The left-most column shows the input image, followed by 4 personalized generations for each subject.