ZipLoRA: Any Subject in Any Style by Effectively Merging LoRAs

Viraj Shah[1,2], Nataniel Ruiz[1], Forrester Cole[1], Erika Lu[1], Svetlana Lazebnik[2], Yuanzhen Li[1], and Varun Jampani[1]

[1] Google Research
[2] UIUC

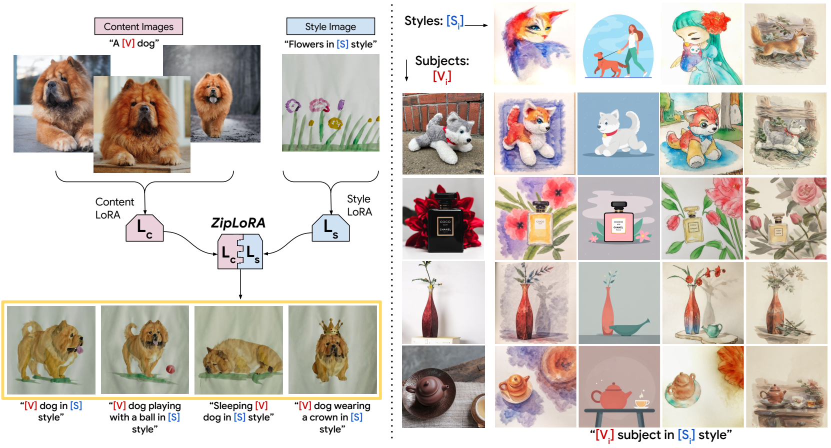


Fig. 1: By effectively merging independently trained style and content LoRAs, our proposed method **ZipLoRA** is able to generate *any user-provided subject in any user-provided style*, providing unprecedented control over personalized creations using diffusion models.

Abstract. Methods for finetuning generative models for concept-driven personalization generally achieve strong results for subject-driven or style-driven generation. Recently, low-rank adaptations (LoRA) have been proposed as a parameter-efficient way of achieving concept-driven personalization. While recent work explores the combination of separate LoRAs to achieve joint generation of learned styles and subjects, existing techniques do not reliably address the problem, so that either subject fidelity or style fidelity are compromised. We propose **ZipLoRA**, a method to cheaply and effectively merge independently trained style and subject LoRAs in order to achieve generation of **any user-provided subject in any user-provided style**. Experiments on a wide range of subject and style combinations show that ZipLoRA can generate compelling results with meaningful improvements over baselines in subject and style fidelity while preserving the ability to recontextualize.

Keywords: Image Stylization · Diffusion Models · LoRA Models

1 Introduction

Recently, diffusion models [13, 30, 36] have allowed for impressive image generation quality with their excellent understanding of diverse artistic concepts and enhanced controllability due to multi-modal conditioning support (with text being the most popular mode). The usability and flexibility of generative models has further progressed with a wide variety of personalization approaches, such as DreamBooth [31] and StyleDrop [35]. These approaches fine-tune a base diffusion model on the images of a specific concept to produce novel renditions in various contexts. Such concepts can be a specific object, person, or artistic style.

While personalization methods have been used for subjects and styles independently, a key unsolved problem is to generate a specific user-provided *subject* in a specific user-provided *style*. For example, an artist may wish to render a specific person in their personal style, learned through examples of their own work. A user may wish to generate images of their child's favorite plush toy, in the style of the child's watercolor paintings. Moreover, if this is achieved two problems are simultaneously solved: (1) the task of representing any given subject in any style, and (2) the problem of controlling diffusion models through images rather than text, which can be imprecise and unsuitable for certain generation tasks. Finally, we can imagine a large-scale application of such a tool, where a bank of independently learned styles and subjects are shared and stored online. The task of arbitrarily rendering *any subject in any style* is an open research problem that we seek to address.

A pitfall of recent personalization methods is that many finetune all of the parameters of a large base model, which can be costly. Parameter Efficient Fine-Tuning (PEFT) approaches allow for fine-tuning models for concept-driven personalization with much lower memory and storage budgets. Among the various PEFT approaches, Low Rank Adaptation (LoRA) [14] has emerged as a favored method for researchers and practitioners alike due to its versatility. LoRA learns low-rank factorized weight matrices for the attention layers (these learned weights are themselves commonly referred to as "LoRAs"). By combining LoRA and algorithms such as DreamBooth [31], the learned subject-specific LoRA weights enable the model to generate the subject with semantic variations.

With the growing popularity of LoRA personalization, there have been attempts to merge LoRA weights, specifically by performing a linear combination of subject and style LoRAs, with variable coefficients [32]. This allows for a control over the "strength" of each LoRA, and users sometimes are able, through careful grid search and subjective human evaluation, to find a combination that allows for accurate portrayal of the subject under the specific style. This method lacks robustness across style and subject combinations, and is also incredibly time consuming.

In this work, we propose *ZipLoRA*, a simple yet effective method to generate any subject in any style by cheaply merging independently trained LoRAs for subject and style. Note that since we aim to achieve custom stylization of a given subject, we focus specifically on merging two LoRAs (one for subject and one for style). Our approach works consistently on a wide variety of subject and style LoRAs without enforcing any restriction on the way these are trained. This allows users and artists to easily combine publicly available subject and style LoRAs of their choice. ZipLoRA is hyperparameter-free, i.e. it does not require manual tuning of any hyperparameters or merger weights.

Our approach is based on two important observations. **(1)** LoRA weights for different layers ΔW_i (where i denotes the layer) are sparse. *i.e.*, most of the elements in ΔW_i have very small magnitude, and have little effect on generation quality and fidelity. **(2)** Columns of the weight matrices of two independently trained LoRAs may have varying levels of "alignment" between each other, as measured by cosine similarity, for example. We find that directly summing columns that are highly aligned degrades performance of the merged model.

Based on these observations, we hypothesize that a method that operates akin to a zipper, aiming to reduce the quantity of similar-direction sums while preserving the content and style generation properties of the original LoRAs will yield more robust, higher-quality merges. Much like a zipper seamlessly joins two sides of a fabric, our proposed optimization-based approach finds a disjoint set of merger coefficients for blending the subject and style LoRAs, ensuring that the merge adeptly captures both subject and style. Our optimization process is lightweight and significantly improves the merging performance on challenging content-style combinations, where the two LoRAs are highly aligned.

While our approach is independent of the model architecture, we further observe that the recently released Stable Diffusion XL (SDXL) model [29] exhibits strong style learning properties, comparable to results shown by StyleDrop [35] on Muse [2]. Specifically, unlike previous versions of Stable Diffusion [30], SDXL is able to learn styles using just a single exemplar image by following a DreamBooth protocol [31] without any human feedback. This property makes our method particularly effective when applied to SDXL. We summarize our contributions as follows:

- We demonstrate some key observations about current text-to-image diffusion models and personalization methods, particularly in relation to style personalization. We further examine the sparsity of concept-personalized LoRA weight matrix coefficients and the prevalence and deleterious effect of highly aligned columns for LoRA matrices.
- Using these insights we propose **ZipLoRA**, a simple optimization method that allows for effective merging of independently trained style and subject LoRAs to allow for the generation of *any subject in any style.*
- We demonstrate the effectiveness of our approach on a variety of image stylization tasks, including content-style transfer and recontextualization. We also demonstrate that ZipLoRA outperforms existing methods of merging LoRAs as well as other baseline approaches.

2 Related Work

Image Stylization. Image-based style transfer is an area of research dating back at least 20 years [5, 12]. Great advances in arbitrary style transfer was achieved by the convolutional neural network-based approaches [9,15,17,24,28]. Generative models such as GANs [18–20] can also be used as a prior for image stylization tasks [1,26,37]. Many recent GAN-based approaches achieve successful one-shot stylizations [3,7,23,25,27,34,38,40–42] by fine-tuning a pre-trained GAN for a given reference style. However, these methods are limited to images from only a single domain (such as faces). Further, most existing GANs do not provide any direct, text-based control over the semantics of the output, thus they cannot produce the reference subject in novel contexts. Methods such as [8,16,22] attempt to modulate the style of the content image using the text description, however, they do not support a style reference image like our approach, and do not provide re-contextualization capability. Compared to older generative models, diffusion models [13,30,36] offer superior generation quality and text-based control; however, to date, it has been difficult to use them for one-shot stylization driven by image examples. Ours is one of the first works demonstrating the use of diffusion models for high-quality example-based stylization combined with an ability to re-contextualize to diverse scenarios.

Fine-tuning of Diffusion Models for Custom Generation. In the evolving field of text-to-image (T2I) model personalization, recent studies have introduced various methods to fine-tune large-scale T2I diffusion models for depicting specific subjects based on textual descriptions. Techniques like Textual Inversion [6] focus on learning text embeddings, while DreamBooth [31] fine-tunes the entire T2I model for better subject representation. Later methods aim to optimize specific parts of the networks [11,21]. Additionally, techniques like LoRA [14] and StyleDrop [35] concentrate on optimizing low-rank approximations and a small subset of weights, respectively, for style personalization. DreamArtist [4] introduces a novel one-shot personalization method using a positive-negative prompt tuning strategy. While these fine-tuning approaches yield high-quality results, they typically are limited to learning only one concept (either subject or style). One exception is Custom Diffusion [21], which attempts to learn multiple concepts simultaneously. However, Custom Diffusion requires expensive joint training from scratch and still yields inferior results when used for stylization as it fails to disentangle the style from the subject.

Combining LoRAs. Combining different LoRAs remain under-explored in the literature particularly from the point of view of fusing style and the subject concepts. Ryu [32] shows a method to combine independently trained LoRAs by weighed arithmetic summation. In [10], authors discuss fusing multiple concept LoRAs using gradient fusion strategy, however, it is an expensive method that requires retraining the entire model. Further, since it uses a custom LoRA variant referred to as ED-LoRA, it lacks the flexibility to combine freely available pre-trained LoRAs. It also relies on regional prompting that uses different prompts for different regions of the image – a trick that is unsuitable for subject-style merge since the style cannot be localized to any one location in the image. A concurrent work discusses a strategy to obtain Mixture of Experts by combining multiple LoRAs using a gating function [39]. However, it focuses only on the ability to generate the individual concepts separately, and does not consider the

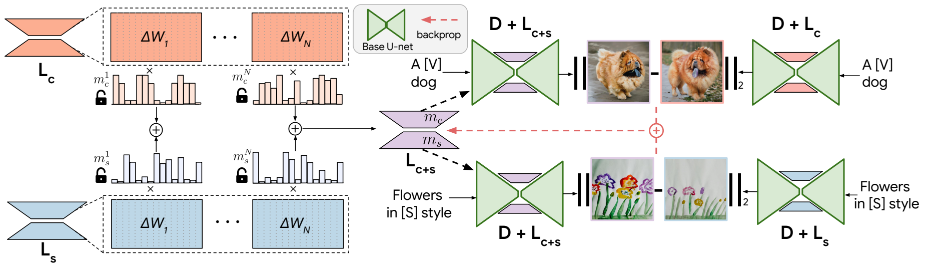


Fig. 2: Overview of ZipLoRA. Our method learns mixing coefficients for each column of ΔW_i for both style and subject LoRAs. It does so by **(1)** minimizing the difference between subject/style images generated by the mixed LoRA and original subject/style LoRA models, while **(2)** minimizing the cosine similarity between the columns of content and style LoRAs. In essence, the zipped LoRA tries to conserve the subject and style properties of each individual LoRA, while minimizing signal interference of both LoRAs.

problem of combined generation, *i.e.* generating multiple different concepts (such as object and style) together in a single image.

3 Methods

3.1 Background

Diffusion Models [13, 30, 36] are state-of-the-art generative models known for their high-quality, photorealistic image synthesis. Their training comprises two phases: a forward process, where an image transitions into a Gaussian noise through incremental Gaussian noise addition, and a reverse process, reconstructing the original data from the noise. The reverse process is typically learnt using an U-net with text conditioning support enabling text-to-image generation at the time of inference. In our work, we focus on widely used latent diffusion model [30] which learns the diffusion process in the latent space instead of image space. In particular, we use Stable Diffusion XL v1 [29] for all our experiments.

LoRA Fine-tuning. LoRA (Low-Rank Adaptation) is a method for efficient adaptation of Large Language and Vision Models to a new downstream task [14, 32]. The key concept of LoRA is that the weight updates ΔW to the base model weights $W_0 \in \mathbb{R}^{m \times n}$ during fine-tuning have a “low intrinsic rank," thus the update ΔW can be decomposed into two low-rank matrices $B \in \mathbb{R}^{m \times r}$ and $A \in \mathbb{R}^{r \times n}$ for efficient parameterization with $\Delta W = BA$. Here, r represents the intrinsic rank of ΔW with $r << min(m, n)$. During training, only A and B are updated to find suitable $\Delta W = BA$, while keeping W_0 constant. For inference, the updated weight matrix W can be obtained as $W = W_0 + BA$. Due to its efficiency, LoRA is widely used for fine-tuning open-sourced diffusion models.

3.2 Problem Setup

In this work, we aim to produce accurate renditions of a custom object in a given reference style by merging LoRA weights obtained by separately fine-tuning a given text-to-image diffusion model on a few reference images of the object/style.

We start with a base diffusion model represented as D with pre-trained weights $W_0^{(i)}$ with i as layer index. One can adapt the base model D to any given concept by simply adding the corresponding set of LoRA weights $L_x\{\Delta W_x^{(i)}\}$ to the model weights. We represent it as: $D_{L_x} = D \oplus L_x = W_0 + \Delta W_x$. We drop the superscript (i) for simplicity since our operations are applied over all the LoRA-enabled weight matrices of our base model D.

We are given two independently trained set of LoRAs $L_c = \{\Delta W_c^{(i)}\}$ and $L_s = \{\Delta W_s^{(i)}\}$ for our base model D, and we aim to find a merged LoRA $L_m = \{\Delta W_m^{(i)}\} = \text{Merge}(L_c, L_s)$ that can combine the effects of both the individual LoRAs in order to stylize the given object in a desired reference style.

Direct Merge. LoRA is popularly used as a plug-and-play module on top of the base model, thus a most common way to combine multiple LoRAs is a simple linear combination [32]:

$$L_m = L_c + L_s \implies \Delta W_m = w_c \cdot \Delta W_c + w_s \cdot \Delta W_s, \tag{1}$$

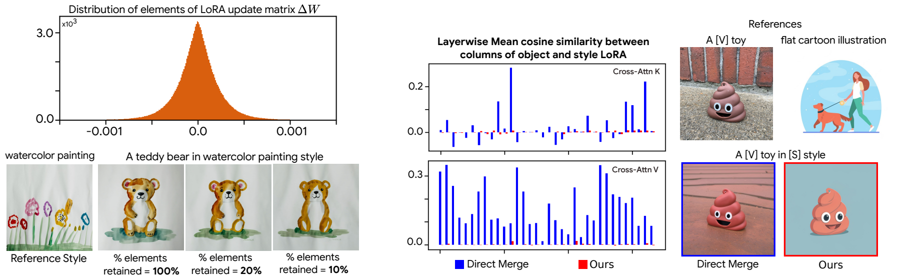
where w_c and w_s are coefficients of content and style LoRAs, respectively, which allow for a control over the "strength" of each LoRA. For a given subject and style LoRA, one may be able to find a particular combination of w_c and w_s that allows for accurate stylization through careful grid search and subjective human evaluation, but this method is not robust and very time consuming. To this end, we propose a hyperparameter-free approach that does not require this onerous process.

3.3 ZipLoRA

Our approach builds on two interesting insights:

(1) LoRA update matrices are sparse. We observe that the update matrices ΔW for different LoRA layers are sparse, *i.e.*, most of the elements in ΔW have a magnitude very close to zero, and thus have little impact on the output of the fine-tuned model. For each layer, we can sort all the elements by their magnitude and zero out the lowest up to a certain percentile. We depict the distribution of elements of $\Delta W_i^{m \times n}$ in Fig. 3a, along with samples generated after zeroing out 80% and 90% of the lowest-magnitude elements of weight update matrix ΔW for all the layers. As can be seen, the model performance is unaffected even when 90% of the elements are thrown away. This observation follows from the fact that the rank of ΔW is very small by design, thus the information contained in most columns of ΔW is redundant.

(2) Highly aligned LoRA weights merge poorly. Columns of the weight matrices of two independently trained LoRAs may contain information that is not disentangled, *i.e.*, the cosine similarity between them can be non-zero. We



(a) LoRA weight matrices are sparse. **(b) Highly aligned LoRA weights merge poorly.**

Fig. 3: Key insights of our approach: **(a)** Most of the elements in ΔW have a magnitude very close to zero, and can be conveniently thrown away without affecting the generation quality of the fine-tuned model. **(b)** When LoRA weight columns are highly aligned, a direct merge obtains subpar results. Instead, our approach minimizes the mean cosine similarity between the columns of the LoRA updates across the layers.

observe that the extent of alignment between the columns of LoRA weights plays a significant role in determining the quality of resulting merge: if we directly add the columns with non-zero cosine similarity to each other, it leads to superimposition of their information about the individual concepts, resulting in the loss of the ability of the merged model to synthesize input concepts accurately. We further observe that such loss of information is avoided when the columns are orthogonal to each other with cosine similarity equal to zero.

Note that each weight matrix represents a linear transformation defined by its columns, so it is intuitive that the merger would retain the information available in these columns only when the columns that are being added are orthogonal to each other. For most content-style LoRA pairs the cosine similarities are non-zero, resulting in signal interference when they are added directly. In Fig. 3b we show the mean cosine similarity values for each layer of the last U-net block for a particular content-style pair before and after applying ZipLoRA. One can see high non-zero cosine similarity values for the direct merge which results in poor stylization quality. On the other hand, ZipLoRA reduces the similarity values significantly to achieve a superior result.

To prevent signal interference during the merger, we multiply each column with a learnable coefficient such that the orthogonality between the columns can be achieved. The fact that LoRA updates are sparse allows us to neglect certain columns from each LoRA, thus facilitating the task of minimizing interference. As shown in Fig. 2, we introduce a set of merger coefficient vectors m_c and m_s for each LoRA layer of the content and style LoRAs, respectively:

$$\begin{aligned} L_m &= \text{Merge}(L_c, L_s, m_c, m_s) \\ \implies \Delta W_m &= m_c \otimes \Delta W_c + m_s \otimes W_s, \end{aligned} \tag{2}$$

where $\otimes$ represents element-wise multiplication between ΔW and broadcasted merger coefficient vector m such that j^{th} column of ΔW gets multiplied with j^{th} element of m. The dimensionalities of m_c and m_s are equal to the number of columns in corresponding $\varDelta W$, thus each element of the merger coefficient vector represents the contribution of the corresponding column of the LoRA matrix $\varDelta W$ to the final merge.

Our ZipLoRA approach has two goals: **(1)** to minimize the interference between content and style LoRAs, defined by the cosine similarity between the columns of content and style LoRAs while **(2)** conserving the capability of the merged LoRA to generate the reference subject and style independently by minimizing the difference between subject/style images generated by the mixed LoRA and original subject/style LoRAs. To ensure that the columns that are merged with each other minimize signal interference, our proposed loss seeks to minimize the alignment between the merge vectors m_c and m_s of each layer. Meanwhile, we wish to ensure that the original behavior of both the style and the content LoRAs is preserved in the merged model. Therefore, as depicted in Fig. 2, we formulate an optimization problem with following loss function:

$$\begin{aligned}\mathcal{L}_{merge} =&\|(D \oplus L_m)(x_c, p_c) - (D \oplus L_c)(x_c, p_c)\|_2 \\ &+\|(D \oplus L_m)(x_s, p_s) - (D \oplus L_s)(x_s, p_s)\|_2 \\ &+\lambda \sum_i |m_c^{(i)} \cdot m_s^{(i)}|,\end{aligned} \tag{3}$$

where the merged model L_m is calculated using m_c and m_s as per Eq. 2; (x_c, x_s) and (p_c, p_s) are noisy latents and text conditioning prompts for content and style references respectively, and λ is an appropriate multiplier for the cosine-similarity loss term. Note that the first two terms ensure that the merged model retains the ability to generate individual style and content, while the third term enforces an orthogonality constraint between the columns of the individual LoRA weights. Importantly, we keep the weights of the base model and the individual LoRAs frozen, and update only the merger coefficient vectors. As seen in the next section, such a simple optimization method is effective in producing strong stylization of custom subjects. Further, ZipLoRA requires only 100 gradient updates which is 10× lower compared to joint training approaches.

4 Experiments

Datasets. We choose a diverse set of content images from the DreamBooth dataset [31], which provides 30 image sets each containing 4-5 images of a given subject. Similarly, a diverse set of style reference images is selected from the data provided by authors of StyleDrop [35]. We use only a single image for each style. The attribution and licence information for all the content and style images used are available in the DreamBooth and StyleDrop manuscripts/websites, and we also include them in the supplementary material.
Experimental Setup. We perform all our experiments using the SDXL v1.0 [29] base model. We use DreamBooth fine-tuning with LoRA of rank 64 for obtaining all the style and content LoRAs. We update the LoRA weights using Adam optimizer for 1000 steps with batch size of 1 and learning rate of 0.00005. We keep

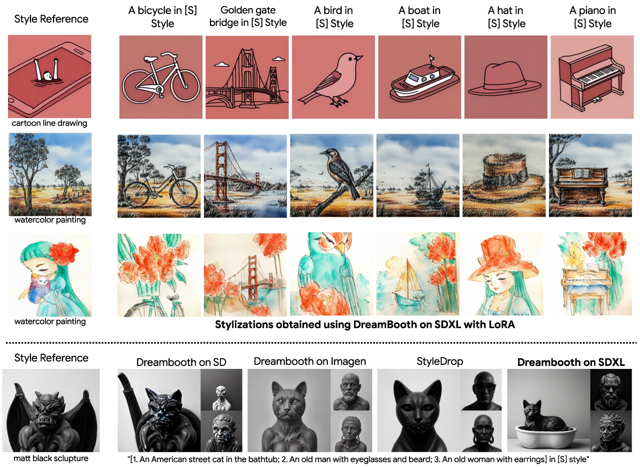


Fig. 4: Style Learning using DreamBooth on SDXL. Top: SDXL model learns to produce stylized outputs when fine-tuned on a single example of a reference style using LoRA with a DreamBooth objective. Bottom: The stylizations produced by fine-tuned SDXL model are superior to those of other models. Note that unlike StyleDrop, SDXL DreamBooth fine-tuning does not require human feedback.

the text encoders of SDXL frozen during the LoRA fine-tuning. For ZipLoRA, we use $\lambda = 0.01$ in Eq. 3 for all our experiments, and run the optimization until cosine similarity drops to zero with a maximum number of gradient updates set to 100. We plan to release the implementation of our method in future. To obtain qualitative and quantitative comparisons with existing methods, we use their official open-source implementations except for StyleDrop [35]. Since the official code and the model for StyleDrop is not available publicly, we obtain its results by contacting the authors.

4.1 Style-tuning behavior of SDXL model

As discussed in Sec. 3, we observe, surprisingly, that a pre-trained SDXL model exhibits strong style learning when fine-tuned on only one reference style image. We show style-tuning results on SDXL model in Fig. 4. For each reference image, we apply LoRA fine-tuning of SDXL model using DreamBooth objective with LoRA rank$= 64$. For fine-tuning, we follow a similar prompt formation as provided in StyleDrop: “an <object> in the <style description> style". Once fine-tuned, SDXL is able to represent diverse set of concepts in the reference style by capturing the nuances of painting style, lighting, colors, and geometry accurately. The question of why this model exhibits this strong style learning performance, as opposed to the lesser performance of previous Stable Diffusion versions [30] (or Imagen [33]) is left open and can have many answers including training data, model architecture and training schemes.

We also provide comparisons with StyleDrop on Muse [2], DreamBooth on Imagen, and DreamBooth on Stable Diffusion (SDv1.5) in Fig. 4. We observe that SDXL style-tuning performs significantly better than the competing methods. Note that StyleDrop requires iterative training with human feedback whereas SDXL style-tuning does not. This behavior of SDXL makes it the perfect candidate for investigating the merging of style LoRAs with subject LoRAs to achieve personalized stylizations. Thus, we choose to use it as a base model for all of our experiments.

4.2 Personalized Stylizations

To start with, we obtain the style LoRAs following the style-tuning on SDXL as described in Sec. 4.1, and obtain object LoRAs by applying DreamBooth fine-tuning on the subject references. Fig. 1 and Fig. 5 show the results of our approach for combining various style and content LoRAs. Our method succeeds at both preserving the identity of the reference subject and capturing the unique characteristics of the reference style.

We also present qualitative comparisons with other approaches in Fig. 5. As a baseline, we compare with the direct arithmetic merge [32] obtained through Eq. 1 with w_c and w_s set to 1. Such direct addition results in loss of information captured in each LoRA and produces inferior results with distorted object and/or style.

We additionally compare our method with joint training of subject and style using a multi-subject variant of DreamBooth with multiple rare unique identifiers. As shown, joint training fails to learn the disentanglement between object and style and produces poor results. It also is the least flexible method since it does not allow the use of pre-trained LoRAs, neither can it be used as a style-only or content-only LoRA. Further, it requires 10× as many training steps as ZipLoRA.

StyleDrop [35] proposes a StyleDrop+DreamBooth approach for achieving personalized stylizations, where a StyleDrop method is applied on a DreamBooth model fine-tuned on the reference object. Our comparisons show that its performance is not ideal, considering the high compute cost and human feedback requirements. It also requires adjusting the object and style model weights w_c and w_s similar to the direct merge in order to produce reasonable outputs, while our method is free from any such hyperparameter tuning.

Further, we compare our method with recent multi-concept generation approaches Mix of Show [10] and Custom Diffusion [21]. Our results reveal that both the methods perform inferior to ZipLoRA. Mix of show relies on region-aware prompting that requires spatial disentanglement between the individual concepts, thus performs poorly for subject-style merge since the style is usually spread across the entire image. Moreover, it uses a custom LoRA model referred as ED-LoRA thus requires training from scratch for each individual concept. Custom Diffusion learns unique text tokens for each concept which does not work reliably when it comes to combining a style with a subject.

Quantitative results. We conduct user studies for a quantitative comparison of our method with existing approaches. In our study, each participant is shown a reference subject and a reference style along with outputs of two methods being

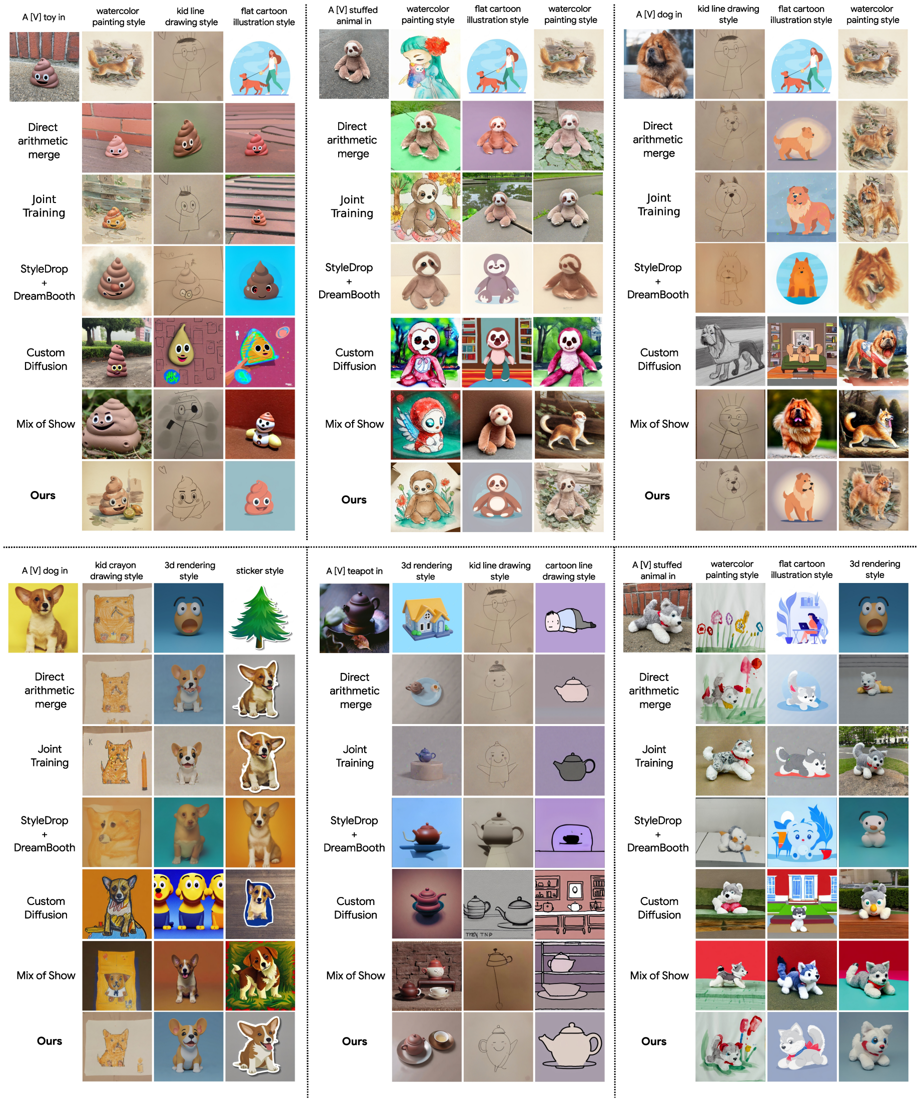


Fig. 5: Qualitative Comparison: We compare samples from our method (Ours), versus direct arithmetic merge [32], joint training, StyleDrop with DreamBooth [35], Mix of Show [10], and Custom Diffusion [21]. We observe that our method achieves strong style and subject fidelity that surpasses competing methods. We provide additional results in Supplementary.

Table 1: **User Preference Study**. We compare the user preference of accurate stylization and subject fidelity between our approach and competing methods. Users generally prefer our approach.

% **Preference for ZipLoRA over:**				
Direct Merge	Joint Training	StyleDrop+DreamBooth	Mix of Show	Custom Diffusion
82.7%	71.1%	68.0%	87.3%	88.1%

Table 2: **Image-alignment and Text-alignment Scores.** We compare cosine similarities between CLIP (for style and text) and DINO features (for subject) of the output and reference style, subject, and prompt respectively. ZipLoRA provides superior subject and text fidelity while also maintaining the style-alignment.

	ZipLoRA	Joint Training	Direct Merge	StyleDrop + DreamBooth	Mix of Show	Custom Diffusion
Style-alignment	0.699	0.680	0.702	0.646	0.635	0.616
Subject-alignment	0.420	0.378	0.357	0.394	0.374	0.346
Text-alignment	0.303	0.296	0.275	0.263	0.251	0.262

compared, in a random order, and asked which output best depicts the reference style while preserving the reference subject fidelity. We conducted separate user studies for ZipLoRA vs. each of the five competing approaches, and received 360 responses across 45 users for each case. We show the results in Tab. 1. As we can see, ZipLoRA receives higher user preference in all three cases owing to its high-quality stylization while preserving subject integrity.

Following DreamBooth [31], we also provide comparisons using image-alignment and text-alignment scores in Tab. 2. We employ three metrics: for style-alignment, we use CLIP-I scores of image embeddings of output and the style reference; for subject-alignment, we employ DINO features for the output and the reference subject; and for text-alignment, we use CLIP-T embeddings of the output and the text prompt. In all three cases, we use cosine-similarity as the metric and calculate averages over 4 subjects in 8 styles each. ZipLoRA results in competitive style-alignment scores as compared to joint training and direct merge, while achieving significantly better scores for subject-alignment. This highlights ZipLoRA's superiority in maintaining the subject fidelity. ZipLoRA also outperforms the existing methods in text-alignment, implying that it preserves the text-to-image generation capability, and also expresses the designated style and

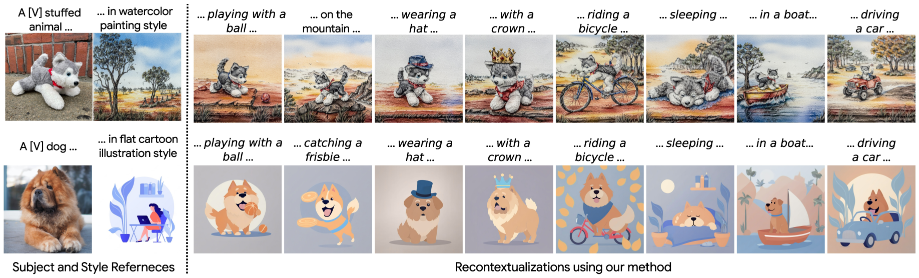
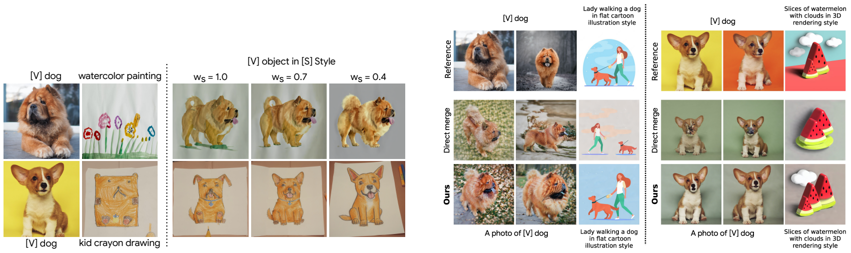


Fig. 6: Our method successfully re-contextualizes the reference subject while preserving the stylization in the given style.


[V] dog
watercolor painting
[V] object in [S] Style
w_S = 1.0
w_S = 0.7
w_S = 0.4
[V] dog
kid crayon drawing
[V] dog
Lady walking a dog in flat cartoon illustration style
Reference
Direct merge
Ours
A photo of [V] dog
Lady walking a dog in flat cartoon illustration style
[V] dog
Slices of watermelon with clouds in 3D rendering style
Reference
Direct merge
Ours
A photo of [V] dog
Slices of watermelon with clouds in 3D rendering style

(a) Style Controllability of ZipLoRA **(b) Retention of the individual concepts**

Fig. 7: (a) Our method works out-of-the-box at achieving good subject and style personalization. Nevertheless, varying the merging weights w_s allows for controlling the extent of stylization. **(b)** Our method does not lose the ability to generate individual concepts, unlike the direct merge approach.

subject better (since these are also part of the text prompt). One should note that these metrics are not perfect, particularly when it comes to measuring style alignment, since they lack the ability to capture subtle stylistic details, and are entangled with semantic properties of images, such as the overall content.

Ability to re-contextualize. The merged ZipLoRA model can recontextualize reference objects in diverse contexts and with semantic modifications while maintaining stylization quality. As shown in Fig. 6, our method preserves the base model's text-to-image generation capabilities while accurately stylizing the entire image in the reference style. Such ability is highly valuable in various artistic use cases that requires controlling contexts, subject identities, and styles.

Controlling the extent of stylization. Our optimization-based method directly provides a scalar weight value for each column of the LoRA update, thus eliminating a need for tuning and adjustments for obtaining reasonable results. However, we can still allow the strength of object and style content to be varied for added controllability. One can attenuate the style layer weights by multiplying them with an additional scalar multiplier w_s to limit the contribution of the style in the final output. As shown in Fig. 7a, this allows for a smooth control over the extent of stylization as w_s varies between 0 to 1.

Ability to produce the reference object and the style. Apart from producing accurate stylizations, an ideal LoRA merge should also preserve the ability to generate individual object and style correctly. This way, a merged LoRA model can also be used as a replacement of both the individual LoRAs. As shown in Fig. 7b, our approach retains the original behavior of both the models and can accurately generate specific structural and stylistic elements of each constituent LoRA, while direct merge fails.

Limitations/failure cases. For some style reference images, instead of capturing just the style, SDXL style-tuning incorrectly captures the subject as well. ZipLoRA fails to disentangle such styles further, thus the content of style reference can leak into the stylization outputs. For example, as shown in Fig. 8, SDXL style-tuning fails to disentangle the cliff from the watercolor painting style, and ZipLoRA ends up producing the cliff in the background in all the stylizations.

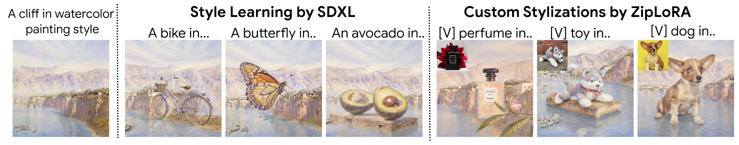


Fig. 8: Failure Cases. For a few styles, ZipLoRA fails to separate the content of the style reference from its style, resulting into the leakage of the content (cliff in this case) in stylization outputs.

Comparisons of runtime/storage. ZipLoRA offers improved efficiency, exhibiting lower storage footprints, reduced computational demands, and faster runtimes. ZipLoRA requires only 100 gradient updates which is 10× less than Joint Training (JT), Custom Diffusion (CD), and Mix of Show (MoS). ZipLoRA's runtime is 560 seconds while JT and CD takes 3540s and 3890s respectively. For MoS, to achieve a successful merger, one first needs to obtain ED-LoRAs for each individual concept, thus the total runtime for MoS is 4980s (1600s each for training ED-LoRAs + 1780s for merging them). All runtimes are calculated on single NVIDIA A100. ZipLoRA updates only the merger coefficient vectors m_c, m_s while keeping the LoRA weights frozen, thus has only 1.6M trainable parameters as opposed to 180M in the case of the competing methods, reducing the GPU memory requirements from 38GB to 21GB. For storage, ZipLoRA needs to store just the merger coefficient vectors m_c, m_s requiring only 6.5MB of storage, while the LoRA resulting from other methods requires 360MB.

Performance on Stable Diffusion (SDv1.5). As discussed in Sec. 4.1 & Fig. 4, LoRA fine-tuning on earlier version of Stable Diffusion (SDv1.5) fails to capture the stylization, thus the performance of ZipLoRA on SDv1.5 becomes limited by the stylization ability of the underlying style LoRA. Our successful style-tuning of SDXL is key observation that led us to adopt it as the base model. That being said, our observations about sparsity and alignment of LoRA weights remain valid for other models of stable diffusion family, and even on SDv1.5, ZipLoRA outperforms competing methods (Direct Merge, Joint Training, Custom Diffusion, and Mix of Show) by achieving better stylizations with improved subject and style fidelity. We provide comparison figure and quantitative results on SDv1.5 in Supplementary.

5 Conclusion and Future Work

In this work, we have introduced **ZipLoRA**, a novel method for seamlessly merging independently trained style and subject LoRAs. Our approach unlocks the ability to generate **any subject in any style** using contemporary diffusion models like SDXL. By leveraging key insights about pre-trained LoRA weights, we surpass existing methods for this task. ZipLoRA offers a streamlined, cheap, and hyperparameter-free solution for simultaneous subject and style personalization, unlocking a new level of creative controllability for diffusion models. While ZipLoRA focuses on merging a pair of a subject and a style LoRA, combining more than two subject/style LoRAs can be considered as a future work.

Acknowledgements. We thank Prafull Sharma, Meera Hahn, Jason Baldridge, and Dilip Krishnan for helpful discussions and suggestions. We also thank Kihyuk Sohn for helping with the generation of StyleDrop results.

References

1. Bora, A., Jalal, A., Price, E., Dimakis, A.G.: Compressed sensing using generative models. In: ICML (2017)
2. 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)
3. Chong, M.J., Forsyth, D.A.: Jojogan: One shot face stylization. CoRR **abs/2112.11641** (2021), https://arxiv.org/abs/2112.11641
4. Dong, Z., Wei, P., Lin, L.: Dreamartist: Towards controllable one-shot text-to-image generation via positive-negative prompt-tuning (2023)
5. Efros, A.A., Freeman, W.T.: Image quilting for texture synthesis and transfer. In: Proceedings of the 28th annual conference on Computer graphics and interactive techniques. pp. 341–346 (2001)
6. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion (2022). https://doi.org/10.48550/ARXIV.2208.01618, https://arxiv.org/abs/2208.01618
7. Gal, R., Patashnik, O., Maron, H., Chechik, G., Cohen-Or, D.: Stylegan-nada: Clip-guided domain adaptation of image generators. ArXiv **abs/2108.00946** (2021)
8. Gal, R., Patashnik, O., Maron, H., Chechik, G., Cohen-Or, D.: Stylegan-nada: Clip-guided domain adaptation of image generators (2021)
9. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2414–2423 (2016)
10. Gu, Y., Wang, X., Wu, J.Z., Shi, Y., Yunpeng, C., Fan, Z., Xiao, W., Zhao, R., Chang, S., Wu, W., Ge, Y., Ying, S., Shou, M.Z.: Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. arXiv preprint arXiv:2305.18292 (2023)
11. Han, L., Li, Y., Zhang, H., Milanfar, P., Metaxas, D., Yang, F.: Svdiff: Compact parameter space for diffusion fine-tuning. arXiv preprint arXiv:2303.11305 (2023)
12. Hertzmann, A., Jacobs, C.E., Oliver, N., Curless, B., Salesin, D.: Image analogies. Proceedings of the 28th annual conference on Computer graphics and interactive techniques (2001)
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. ArXiv **abs/2006.11239** (2020), https://api.semanticscholar.org/CorpusID:219955663
14. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=nZeVKeeFYf9
15. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1501–1510 (2017)

16. Jandial, S., Deshmukh, S., Java, A., Shahid, S., Krishnamurthy, B.: Gatha: Relational loss for enhancing text-based style transfer. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) pp. 3546–3551 (2023), `https://api.semanticscholar.org/CorpusID:260919917`
17. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European Conference on Computer Vision (2016)
18. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data. ArXiv **abs/2006.06676** (2020)
19. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 4396–4405 (2019)
20. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 8107–8116 (2020)
21. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: CVPR (2023)
22. Kwon, G., Ye, J.C.: Clipstyler: Image style transfer with a single text condition. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 18041–18050 (2021), `https://api.semanticscholar.org/CorpusID:244773443`
23. Kwon, G., Ye, J.C.: One-shot adaptation of gan in just one clip. ArXiv **abs/2203.09301** (2022)
24. Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., Yang, M.H.: Universal style transfer via feature transforms. In: Advances in Neural Information Processing Systems (2017)
25. Liu, M., Li, Q., Qin, Z., Zhang, G., Wan, P., Zheng, W.: Blendgan: Implicitly gan blending for arbitrary stylized face generation. ArXiv **abs/2110.11728** (2021)
26. Menon, S., Damian, A., Hu, S., Ravi, N., Rudin, C.: Pulse: Self-supervised photo upsampling via latent space exploration of generative models. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2434–2442 (2020)
27. Ojha, U., Li, Y., Lu, J., Efros, A.A., Lee, Y.J., Shechtman, E., Zhang, R.: Few-shot image generation via cross-domain correspondence. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 10738–10747 (2021)
28. Park, D.Y., Lee, K.H.: Arbitrary style transfer with style-attentional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5880–5888 (2019)
29. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Muller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. ArXiv **abs/2307.01952** (2023), `https://api.semanticscholar.org/CorpusID:259341735`
30. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 10674–10685 (2021), `https://api.semanticscholar.org/CorpusID:245335280`
31. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22500–22510 (2023)
32. Ryu, S.: Merging loras. `https://github.com/cloneofsimo/lora`

33. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding. ArXiv **abs/2205.11487** (2022), https://api.semanticscholar.org/CorpusID:248986576
34. Shah, V., Sarkar, A., Anita, S.K., Lazebnik, S.: Multistylegan: Multiple one-shot image stylizations using a single gan. arXiv (2023)
35. Sohn, K., Ruiz, N., Lee, K., Chin, D.C., Blok, I., Chang, H., Barber, J., Jiang, L., Entis, G., Li, Y., et al.: Styledrop: Text-to-image generation in any style. arXiv preprint arXiv:2306.00983 (2023)
36. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. ArXiv **abs/2010.02502** (2020), https://api.semanticscholar.org/CorpusID:222140788
37. Wang, X., Li, Y., Zhang, H., Shan, Y.: Towards real-world blind face restoration with generative facial prior. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 9164–9174 (2021)
38. Wang, Y., Yi, R., Tai, Y., Wang, C., Ma, L.: Ctlgan: Few-shot artistic portraits generation with contrastive transfer learning. ArXiv **abs/2203.08612** (2022)
39. Wu, X., Huang, S., Wei, F.: MoLE: Mixture of loRA experts. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=uWvKBCYh4S
40. Yang, C., Shen, Y., Zhang, Z., Xu, Y., Zhu, J., Wu, Z., Zhou, B.: One-shot generative domain adaptation (2021). https://doi.org/10.48550/ARXIV.2111.09876, https://arxiv.org/abs/2111.09876
41. Yang, S., Jiang, L., Liu, Z., Loy, C.C.: Pastiche master: Exemplar-based high-resolution portrait style transfer (2022). https://doi.org/10.48550/ARXIV.2203.13248, https://arxiv.org/abs/2203.13248
42. Zhu, P., Abdal, R., Femiani, J., Wonka, P.: Mind the gap: Domain gap control for single shot domain adaptation for generative adversarial networks. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=vqGi8Kp0wM