SVDiff: Compact Parameter Space for Diffusion Fine-Tuning

Ligong Han\textsuperscript{1,2}\textsuperscript{*}  Yinxiao Li\textsuperscript{2}  Han Zhang\textsuperscript{2}  Peyman Milanfar\textsuperscript{2}  Dimitris Metaxas\textsuperscript{1}  Feng Yang\textsuperscript{2}
\textsuperscript{1}Rutgers University  \textsuperscript{2}Google Research

Abstract

Diffusion models have achieved remarkable success in text-to-image generation, enabling the creation of high-quality images from text prompts or other modalities. However, existing methods for customizing these models are limited by handling multiple personalized subjects and the risk of overfitting. Moreover, their large number of parameters is inefficient for model storage. In this paper, we propose a novel approach to address these limitations in existing text-to-image diffusion models for personalization. Our method involves fine-tuning the singular values of the weight matrices, leading to a compact and efficient parameter space that reduces the risk of overfitting and language-drifting. We also propose a Cut-Mix-Unmix data-augmentation technique to enhance the quality of multi-subject image generation and a simple text-based image editing framework. Our proposed SVDiff method has a significantly smaller model size compared with existing methods (≈2,200 times fewer parameters compared with vanilla DreamBooth), making it more practical for real-world applications.

1. Introduction

Recent years have witnessed the rapid advancement of diffusion-based text-to-image generative models [19, 53, 20, 47, 51], which have enabled the generation of high-quality images through simple text prompts. These models are capable of generating a wide range of objects, styles, and scenes with remarkable realism and diversity. These models, with their exceptional results, have inspired researchers to investigate various ways to harness their power for image editing [31, 37, 79].

In the pursuit of model personalization and customization, some recent works such as Textual-Inversion [16], DreamBooth [52], and Custom Diffusion [32] have further unleashed the potential of large-scale text-to-image diffusion models. By fine-tuning the parameters of the pre-trained models, these methods allow the diffusion models to be adapted to specific tasks or individual user preferences.

Despite their promising results, there are still some limitations associated with fine-tuning large-scale text-to-image diffusion models. One limitation is the large parameter space, which can lead to overfitting or drifting from the original generalization ability [52]. Another challenge is the difficulty in learning multiple personalized concepts especially when they are of similar categories [32].

To alleviate overfitting, we draw inspiration from the efficient parameter space in the GAN literature [50] and propose a compact yet efficient parameter space, spectral shift, for diffusion model by only fine-tuning the singular values of the weight matrices of the model. This approach is inspired by prior work in GAN adaptation showing that constraining the space of trainable parameters can lead to improved performance on target domain [48, 36, 41, 64]. Comparing with another popular low-rank constraint [22], the spectral shifts utilize the full representation power of the weight matrix while being more compact (\textit{e.g.} 1.7MB for

\*Work done during an internship at Google Research.
StableDiffusion [51, 15, 69], full weight checkpoint consumes 3.66GB of storage). The compact parameter space allows us to combat overfitting and language-drifting issues, especially when prior-preservation loss [52] is not applicable. We demonstrate this use case by presenting a simple DreamBooth-based single-image editing framework.

To further enhance the ability of the model to learn multiple personalized concepts, we propose a simple Cut-Mix-Unmix data-augmentation technique. This technique, together with our proposed spectral shift parameter space, enables us to learn multiple personalized concepts even for semantically similar categories (e.g., a “cat” and a “dog”).

In summary, our main contributions are:

• We present a compact (≈2,200× fewer parameters compared with vanilla DreamBooth [52], measured on StableDiffusion [51]) yet efficient parameter space for diffusion model fine-tuning based on singular-value decomposition of weight kernels.

• We present a text-based single-image editing framework and demonstrate its use case with our proposed spectral shift parameter space.

• We present a generic Cut-Mix-Unmix method for data-augmentation to enhance the ability of the model to learn multiple personalized concepts.

This work opens up new avenues for the efficient and effective fine-tuning large-scale text-to-image diffusion models for personalization and customization. Our proposed method provides a promising starting point for further research in this direction.

2. Related Work

Text-to-image diffusion models Diffusion models [58, 61, 19, 62, 40, 59, 18, 60, 10] have proven to be highly effective in learning data distributions and have shown impressive results in image synthesis, leading to various applications [74, 45, 9, 49, 26, 23, 34, 56, 28, 24, 29]. Recent advancements have also explored transformer-based architectures [67, 44, 6, 7]. In particular, the field of text-guided image synthesis has seen significant growth with the introduction of diffusion models, achieving state-of-the-art results in large-scale text-to-image synthesis tasks [39, 47, 53, 51, 4]. Our main experiments were conducted using StableDiffusion [51], which is a popular variant of latent diffusion models (LDMs) [51] that operates on a latent space of a pre-trained autoencoder to reduce the dimensionality of the data samples, allowing the diffusion model to utilize the well-compressed semantic features and visual patterns learned by the encoder.

Fine-tuning generative models for personalization Recent works have focused on customizing and personalizing text-to-image diffusion models by fine-tuning the text embedding [16], full weights [52], cross-attention layers [32], or adapters [77, 38] using a few personalized images. Other works have also investigated training-free approaches for fast adaptation [17, 73, 11, 27, 55]. The idea of fine-tuning only the singular values of weight matrices was introduced by FSGAN [50] in the GAN literature and further advanced by NaviGAN [13] with an unsupervised method for discovering semantic directions in this compact parameter space. Our method, SVDiff, introduces this concept to the fine-tuning of diffusion models and is designed for few-shot adaptation. A similar approach, LoRA [14], explores low-rank adaptation for text-to-image diffusion fine-tuning, while our proposed SVDiff optimizes all singular values of the weight matrix, leading to an even smaller model checkpoint. Similar idea has also been explored in few-shot segmentation [64].

Diffusion-based image editing Diffusion models have also shown great potential for semantic editing [33, 3, 2, 31, 79, 37, 72, 68, 63, 42, 43, 5, 71]. These methods typically focus on inversion [59] and reconstruction by optimizing the nullspace embedding or overfitting to the given image [79]. Our proposed method, SVDiff, presents a simple DreamBooth-based [52] single-image editing framework that demonstrates the potential of SVDiff in single image editing and mitigating overfitting.

3. Method

3.1. Preliminary

Diffusion models StableDiffusion [51], the model we experiment with, is a variant of latent diffusion models (LDMs) [51]. LDMs transform the input images x into a latent code z through an encoder $\mathcal{E}$, where $z = \mathcal{E}(x)$, and perform the denoising process in the latent space $\mathcal{Z}$. Briefly, a LDM $\mathcal{E}_\theta$ is trained with a denoising objective:

$$E_{z,c,t} \left[ \| \mathcal{E}_\theta(z_c, t, c) - \epsilon \|_2^2 \right],$$

where $(z, c)$ are data-conditioning pairs (image latents and text embeddings), $\epsilon \sim \mathcal{N}(0, I)$, $t \sim \text{Uniform}(1, T)$, and $\theta$ represents the model parameters. We omit $t$ in the following for brevity.

Few-shot adaptation in compact parameter space of GANs The method of FSGAN [50] is based on the Singular Value Decomposition (SVD) technique and proposes an effective way to adapt GANs in few-shot settings. It takes advantage of the SVD to learn a compact update for domain adaptation in the parameter space of a GAN. Specifically, FSGAN reshapes the convolution kernels of a GAN, which are in the form of $W_{\text{tensor}} \in \mathbb{R}^{c_{out} \times c_{in} \times h \times w}$, into 2-D matrices $W$, which are in the form of $W = \text{reshape}(W_{\text{tensor}}) \in \mathbb{R}^{c_{out} \times (c_{in} \times h \times w)}$. FSGAN then performs SVD on these reshaped weight matrices of both
the generator and discriminator of a pretrained GAN and adapts their singular values to a new domain using a standard GAN training objective.

3.2. Compact Parameter Space for Diffusion Fine-tuning

Spectral shifts The core idea of our method is to introduce the concept of spectral shifts from FSGAN [50] to the parameter space of diffusion models. To do so, we first perform Singular Value Decomposition (SVD) on the weight matrices of the pre-trained diffusion model. The weight matrix (obtained from the same reshaping as FSGAN [50] mentioned above) is denoted as $W$ and its SVD is $W = U \Sigma V^T$, where $\Sigma = \text{diag}(\sigma)$ and $\sigma = [\sigma_1, \sigma_2, \ldots]$ are the singular values in descending order. Note that the SVD is a one-time computation and can be cached. This procedure is illustrated in Fig. 2. Such reshaping of the convolution kernels is inspired by viewing them as linear associative memories [8]. The patch-level convolution can be expressed as a matrix multiplication, $f_{out} = W f_{in}$, where $f_{in} \in \mathbb{R}^{(c_i \times h \times w) \times 1}$ is flattened patch feature and $f_{out} \in \mathbb{R}^{c_o \times 1}$ is the output pre-activation feature corresponding to the given patch. Intuitively, the optimization of spectral shifts leverages the fact that the singular vectors correspond to the close-form solutions of the eigenvalue problem [54]: $\max_n \|W n\|^2_s.t. \|n\| = 1$.

Instead of fine-tuning the full weight matrix, we only update the weight matrix by optimizing the spectral shift [13], $\delta$, which is defined as the difference between the singular values of the updated weight matrix and the original weight matrix. The updated weight matrix can be re-assembled by

$$W_\delta = U \Sigma_\delta V^T \text{ with } \Sigma_\delta = \text{diag}(\text{ReLU}(\sigma + \delta)). \quad (2)$$

Training loss The fine-tuning is performed using the same loss function that was used for training the diffusion model, with a weighted prior-preservation loss [52, 8]:

$$\mathcal{L}(\delta) = \mathbb{E}_{z^*, c^*, \epsilon, t}[\|\hat{e}_{\theta_\delta}(z^*_t | c^*) - \epsilon\|^2_2] + \lambda \mathcal{L}_{pr}(\delta) \text{ with } \mathcal{L}_{pr}(\delta) = \mathbb{E}_{z^{pr}, c^{pr}, \epsilon, t}[\|\hat{e}_{\theta_\delta}(z^{pr}_t | c^{pr}) - \epsilon\|^2_2] \quad (3)$$

where $(z^*, c^*)$ represents the target data-conditioning pairs that the model is being adapted to, and $(z^{pr}, c^{pr})$ represents the prior data-conditioning pairs generated by the pre-trained model. This loss function extends the one proposed by Model Rewriting [8] for GANs to the context of diffusion models, with the prior-preservation loss serving as the smoothing term. We set $\lambda = 0$ for single image editing.

Combining spectral shifts Moreover, the individually trained spectral shifts can be combined into a new model to create novel renderings. This can enable applications including interpolation, style mixing, or multi-subject generation (Fig. 8). Here we consider two common strategies, addition and interpolation. To add $\delta_1$ and $\delta_2$ into $\delta'$,

$$\Sigma_{\delta'} = \text{diag}((\text{ReLU}(\sigma + \delta_1 + \delta_2))). \quad (4)$$

For interpolation between two models with $0 \leq \alpha \leq 1$,

$$\Sigma_{\delta'} = \text{diag}((\text{ReLU}(\sigma + \alpha \delta_1 + (1 - \alpha) \delta_2))). \quad (5)$$

This allows for smooth transitions between models and the ability to interpolate between different image styles.

3.3. Cut-Mix-Unmix for Multi-Subject Generation

We discovered that when training the StableDiffusion [51] model with multiple concepts simultaneously (randomly choosing one concept at each data sampling iteration), the model tends to mix their styles when rendering them in one image for difficult compositions or subjects of similar categories [32] (as shown in Fig. 6). To explicitly
guide the model not to mix personalized styles, we propose a simple technique called Cut-Mix-Unmix. By constructing and presenting the model with “correctly” cut-and-mixed image samples (as shown in Fig. 3), we instruct the model to unmix styles. In this method, we manually create CutMix-like [76] image samples and corresponding prompts (e.g. “photo of a [V1] dog on the left and a [V2] sculpture on the right” or “photo of a [V2] sculpture and a [V1] dog” as illustrated in Fig. 3). The prior loss samples are generated in a similar manner. During training, Cut-Mix-Unmix data augmentation is applied with a pre-defined probability (usually set to 0.6). This probability is not set to 1, as doing so would make it challenging for the model to differentiate between subjects. During inference, we use a different prompt from the one used during training, such as “a [V1] dog sitting beside a [V2] sculpture”. However, if the model overfits to the Cut-Mix-Unmix samples, it may generate samples with stitching artifacts even with a different prompt. We found that using negative prompts can sometimes alleviate these artifacts, as detailed in appendix.

We further present an extension to our fine-tuning approach by incorporating an “unmix” regularization on the cross-attention maps. This is motivated by our observation that in fine-tuned models, the dog’s special token attends largely to the panda’s region. To enforce separation between the two subjects, we use MSE on the non-corresponding regions of the cross-attention maps. This loss encourages the dog’s special token to focus solely on the dog and vice versa for the panda. The results of this extension show a significant reduction in stitching artifact. Details of the cross-attention regularization are presented in the appendix.

### 3.4. Single-Image Editing

In this section, we present a framework for single image editing, by fine-tuning a diffusion model with an image-prompt pair. The procedure is outlined in Fig. 4. The desired edits can be obtained at inference time by modifying the prompt. For example, we fine-tune the model with the input image and text description “photo of a crown with a blue diamond and a golden eagle on it”, and at inference time if we want to remove the eagle, we simply sample from the fine-tuned model with text “photo of a crown with a blue diamond on it”. To mitigate overfitting during fine-tuning, we use the spectral shift parameter space instead of full weights, reducing the risk of overfitting and language drifting. The trade-off between faithful reconstruction and editability, as discussed in [35], is acknowledged, and the purpose of employing SVDiff here is to allow more flexible edits rather than exact reconstructions.

For edits that do not require large structural changes (like repose, “standing” → “lying down” or “zoom in”), results can be improved with DDIM inversion [59]. Before sampling, we run DDIM inversion with classifier-free guid-
<table>
<thead>
<tr>
<th>Input Images</th>
<th>SVDiff (Ours)</th>
<th>DreamBooth</th>
<th>LoRA</th>
<th>Custom Diffusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;photo of a 𝑉 dog&quot; (a)</td>
<td>&quot;a 𝑉 dog in a swimming pool&quot;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;… a 𝑉 sculpture&quot; (b)</td>
<td>&quot;a 𝑉 sculpture made of glass&quot;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;… a 𝑉 sculpture&quot; (c)</td>
<td>&quot;a 𝑉 sculpture holding flowers&quot;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;… a 𝑉 plushy&quot; (d)</td>
<td>&quot;a 𝑉 plushy on a skateboard in times square&quot;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;… a 𝑉 building&quot; (e)</td>
<td>&quot;a 𝑉 building in Minecraft style&quot;</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 5. Results for single subject generation. DreamBooth [52] and Custom Diffusion [32] are implemented in StableDiffusion with Diffusers library [69]. Each subfigure consists 3 samples: a large one on the left and 2 small one on the right. The text prompt under input images are used for training and the text prompt under sample images are used for inference. We observe that SVDiff performs similarly as DreamBooth, and preserves subject identities better than Custom Diffusion for row 2, 3, 5.

and ablations. The DDIM [59] sampler with \( \eta = 0 \) is used for all generated samples, unless specified otherwise.

### 4.1. Single-Subject Generation

In this section, we present the results of our proposed SVDiff for customized single-subject generation proposed in DreamBooth [52], which involves fine-tuning the pre-trained text-to-image diffusion model on a single object or concept (using 3-5 images). The original DreamBooth was implemented on Imagen [53] and we conduct our experiments based on its StableDiffusion [51] implementation [75, 69]. We provide visual comparisons of 5 examples in Fig. 5. All baselines were trained for 500 or 1000 steps with batch size 1 (except for Custom Diffusion [32], which used a default batch size of 2), and the best model was selected for fair comparison. As Fig. 5 shows, SVDiff produces similar results to DreamBooth (which fine-tunes the full model weights) despite having a much smaller parameter space. Custom Diffusion, on the other hand, tends to underfit the training images as seen in rows 2, 3, and 5 of Fig. 5. We assess the text and image alignment in Fig. 9. The results show that the performance of SVDiff is similar to that of DreamBooth, while Custom Diffusion tends to underfit as seen from its position in the upper left corner.

### 4.2. Multi-Subject Generation

In this section, we present the multi-subject generation results to illustrate the advantage of our proposed “Cut-Mix-Unmix” data augmentation technique. When enabled, we perform Cut-Mix-Unmix data-augmentation with probability 0.6 in each data sampling iteration and two subjects are randomly selected without replacement. A comparison between using “Cut-Mix-Unmix” (marked as “w/ Cut-Mix-Unmix”) and not using it (marked as “w/o Cut-Mix-Unmix”, performing augmentation with probability 0) are shown in Fig. 6. Each row of images are generated using the same text prompt displayed below the images. Note that the Cut-Mix-Unmix data augmentation technique is generic and can be applied to fine-tuning full weights as well.

To assess the visual quality of images generated using the “Cut-Mix-Unmix” method with either SVD or full weights, we conducted a user study using Amazon MTurk [1] with 400 generated image pairs. The participants were presented with an image pair generated using the same random seed, and were asked to identify the better image by answering the question, “Which image contains both objects from the two input images with a consistent background?” Each image pair was evaluated by 10 different raters, and the aggregated results showed that SVD was favored over full weights 60.9% of the time, with a standard deviation of 6.9%. More details and analysis will be provided in the appendix.

Additionally, we also conducted experiments that involve training on three concepts simultaneously. During training, we still construct Cut-Mix samples with probability 0.6 by randomly sample two subjects. Interestingly, we observe that for concepts that are already semantically well-separated, e.g. “dog/building” or “sculpture/building”, the model can successfully generate desired results even without using Cut-Mix-Unmix. However, it fails to disentangle semantically more similar concepts, e.g. “dog/panda” as shown in Fig. 6-g.

### 4.3. Single Image Editing

In this section, we present results for the single image editing application. As depicted in Fig. 7, each row presents three edits with fine-tuning of both spectral shifts (marked as “Ours”) and full weights (marked as “Full”). The text prompts for the corresponding edited images are given below the images. The aim of this experiment is to demonstrate that regularizing the parameter space with spectral shifts effectively mitigates the language drift issue, as defined in [52] (the model overfits to a single image and loses its ability to generalize and perform desired edits).

As previously discussed, when DDIM inversion is not employed, fine-tuning with spectral shifts can lead to sometimes over-creative results. We show examples and comparisons of editing results with and without DDIM inver-
Figure 6. Results for multi-subject generation. (a-d) show the results of fine-tuning on two subjects and (e-g) show the results of fine-tuning on three subjects. Both full weight (“Full”) fine-tuning and SVDiff (“SVD”) can benefit from the Cut-Mix-Unmix data-augmentation. Without Cut-Mix-Unmix, the model struggles to disentangle subjects of similar categories, as demonstrated in the last two columns of (a,b,c,d,g).
Figure 7. Results for single image editing. SVDiff (“Ours”) enables successful image edits despite slight misalignment with the original image. SVDiff performs desired modifications when full model fine-tuning (“Full”) fails, such as removing an object (2nd edit in (a)), adjusting pose (2nd edit in (c)), or zooming in (3rd edit in (d)). The backgrounds in some cases may be affected, however, the subject of the image remains well-preserved. For ours, we use DDIM inversion [59] for all edits in (a,c,e) and the first edit in (d).

Parameter subsets We explore the fine-tuning of spectral inversion [59] in the appendix. Our results show that DDIM inversion improves the editing quality and alignment with the input image for non-structural edits when using our spectral shift parameter space, but may worsen the results for full weight fine-tuning. For example, in Fig. 7, we use DDIM inversion for the edits in (a,c,e) and the first edit in (d). The second edit in (d) presents an interesting example where our method can actually make the statue hold an apple with its hand. Additionally, our fine-tuning approach still produces the desired edit of an empty room even with DDIM inversion, as seen in the third edit of Fig. 7-a. Overall, we see that SVDiff can still perform desired edits when full model fine-tuning exhibits language drift, i.e. it fails to remove the picture in the second edit of (a), change the pose of the dog in the second edit of (c), and zoom-in view in (d).

4.4. Analysis and Ablation

Due to space limitations, we present parameter subsets, weight combination, interpolation and style mixing analysis in this section and provide further analysis including rank, scaling, and correlation in the appendix.
Figure 8. Effects of combining spectral shifts ($\Sigma_{d^t} = \text{diag}(\text{ReLU}(\sigma + \delta_1 + \delta_2))$) and weight deltas ($W' = W + \Delta W_1 + \Delta W_2$) in one model. The combined model retains individual subject features but may mix styles for similar subjects. The results also suggests that the task arithmetic property [25] of language models also holds in StableDiffusion.

<table>
<thead>
<tr>
<th>Subset</th>
<th>SVDiff Parameters</th>
<th>Storage</th>
<th>Subset</th>
<th>SVDiff Parameters</th>
<th>Storage</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNet</td>
<td>all UNet layers</td>
<td>1404KB</td>
<td>Up-Blocks</td>
<td>up-blocks in UNet</td>
<td>789KB</td>
</tr>
<tr>
<td>UNet-CA</td>
<td>all CrossAttn layers in UNet</td>
<td>194KB</td>
<td>Down-Blocks</td>
<td>down-blocks in UNet</td>
<td>469KB</td>
</tr>
<tr>
<td>UNet-CA-KV</td>
<td>$W^K, W^V$ in CrossAttn in UNet</td>
<td>84.8KB</td>
<td>Mid-Block</td>
<td>mid-blocks in UNet</td>
<td>135KB</td>
</tr>
<tr>
<td>UNet-1D</td>
<td>all 1-D weights in UNet</td>
<td>430KB</td>
<td>Up-CA</td>
<td>CrossAttn in up-blocks</td>
<td>106KB</td>
</tr>
<tr>
<td>UNet-2D</td>
<td>all 2-D weights in UNet</td>
<td>617KB</td>
<td>Down-CA</td>
<td>CrossAttn in down-blocks</td>
<td>70.4KB</td>
</tr>
<tr>
<td>UNet-4D</td>
<td>all 4-D weights in UNet</td>
<td>355KB</td>
<td>Mid-CA</td>
<td>CrossAttn in mid-block</td>
<td>17.7KB</td>
</tr>
</tbody>
</table>

Table 1. Fine-tuning 12 subsets of parameters in UNet, along with their corresponding model sizes.

Figure 9. (a) Correlation of individually learned spectral shifts for different subjects. The cosine similarities between the spectral shifts of two subjects are averaged across all layers and plotted. The diagonal shows average similarities between two runs with different learning rates. (b) Text- and image-alignment for single-subject generation. The generated image is denoted as $x'$. The text-alignment is measured by the CLIP score [46, 12] $\cos(\hat{x}, c)$, and the image-alignment is defined as $1 - \mathcal{L}_{\text{LPIPS}}(\hat{x}, x^*)$ [78].

shifts within a subset of parameters in UNet. We consider 12 distinct subsets for our ablation study, as outlined in Tab. 1. Due to space limitations, we provide the visual samples and text-/image-alignment scores for each subset on 5 subjects in appendix. Our findings are as follows: (1) Optimizing the cross-attention (CA) layers generally results in better preservation of subject identity compared to optimizing key and value projections. (2) Optimizing the up-, down-, or mid-blocks of UNet alone is insufficient to maintain identity, which is why we did not further isolate subsets of each part. However, it appears that the up-blocks exhibit the best preservation of identity. (3) In terms of dimensionality, the 2D weights demonstrate the most influence, and offer better identity preservation than UNet-CA.

Weight combination We analyze the effects of weight combination by Eq. (4). Fig. 8 shows a comparison between combining only spectral shifts (marked in “SVD”) and combining the full weights (marked in “Full”). The
The combined model in both cases retains unique features for individual subjects, but may blend their styles for similar concepts (as seen in (e)). For dissimilar concepts (such as the [V_2] sculpture and [V_3] building in (j)), the models can still produce separate representations of each subject. Interestingly, combining full weight deltas can sometimes result in better preservation of individual concepts, as seen in the clear building feature in (j). We posit that this is due to the fact that SVDiff limits update directions to the eigenvectors, which are identical for different subjects. As a result, summing individually trained spectral shifts tends to create more “interference” than summing full weight deltas.

**Style transfer and mixing** A simple and straightforward way to enable style-mixing is to combine individually trained spectral shifts using Eq. (4) and inference with the combined model. We show visual examples of this strategy in appendix and further explore a more challenging and controllable approach for style-mixing as follows. Inspired by the disentangling property observed in StyleGAN [30], we hypothesize that a similar property applies in our context. Following Extended Textual Inversion (XTI [70]), we conducted a style mixing experiment, as illustrated in Fig. 10. For this experiment, we fine-tuned SVDiff on the UNet-2D subset and employed the geometry information provided by (16, down’, 1) – (8, down’, 0) (as described in XTI, Section 8.1). We observe that our spectral shift parameter space allows us to achieve a similar disentangled style-mixing effect, comparable to the P+ space in XTI.

**Interpolation** Fig. 11 shows the results of weight interpolation for both spectral shifts and full weights. The models are marked as “SVD” and “Full”, respectively. The first two rows of the figure demonstrate interpolating between two different classes, such as “dog” and “sculpture”, using the same abstract class word “thing” for training. Each column shows the sample from α-interpolated models. For spectral shifts (“SVD”), we use Eq. (5) and for full weights, we use \( W' = W + \alpha \Delta W_1 + (1-\alpha) \Delta W_2 = \alpha W_1 + (1-\alpha) W_2 \). The images in each row are generated using the same random seed with the deterministic DDIM sampler [59] (η = 0). As seen from the results, both spectral shift and full weight interpolation are capable of generating intermediate concepts between the two original classes.

### 4.5. Comparison with LoRA

In our comparison of SVDiff and LoRA [14, 22] for single image editing, we find that while LoRA tends to underfit, SVDiff provides a balanced trade-off between faithfulness and realism. Additionally, SVDiff results in a significantly smaller delta checkpoint size, being 1/2 to 1/3 that of LoRA. However, in cases where the model requires extensive fine-tuning or learning of new concepts, LoRA’s flexibility to adjust its capability by changing the rank may be beneficial. Further research is needed to explore the potential benefits of combining these approaches. A comparison can be found in the appendix.

It is noteworthy that, with rank one, the storage and update requirements for the \( W \) matrix of shape \( M \times N \) in SVDiff are min(\( M, N \)) floats, compared to (\( M + N \)) floats for LoRA. This may be useful for amortizing or developing training-free approaches for DreamBooth [52]. Additionally, exploring functional forms [65, 66] of spectral shifts is an interesting avenue for future research.

### 5. Conclusion and Limitation

In conclusion, we have proposed a compact parameter space, spectral shift, for diffusion model fine-tuning. The results of our experiments show that fine-tuning in this parameter space achieves similar or even better results compared to full weight fine-tuning in both single- and multi-subject generation. Our proposed Cut-Mix-Unmix data-augmentation technique also improves the quality of multi-subject generation, making it possible to handle cases where subjects are of similar categories. Additionally, spectral shift serves as a regularization method, enabling new use cases like single image editing.

**Limitations** Our method has certain limitations, including the decrease in performance of Cut-Mix-Unmix as more subjects are added and the possibility of an inadequately-preserved background in single image editing. Despite these limitations, we see great potential in our approach for fine-tuning diffusion models and look forward to exploring its capabilities further in future research, such as combining spectral shifts with LoRA or developing training-free approaches for fast personalizing concepts.

**Acknowledgments** This research has been partially funded by research grants to D. Metaxas through NSF: IUCRC CARTA 1747778, 2235405, 2212301, 1951890, 2003874, and NIH-5R01HL127661.


