ISSN: 2581-4311
DOI: doi.org/10.30967/IJCRSET/Alberto-Hojel/151

URF PUBLISHERS
connect with research world

International Journal of Current Research in Science, Engineering & Technology

https://urfpublishers.com/journal/ijcrset

Vol: 7 & Iss: 4

Review

Lora Diffusion: Zero-Shot Lora Synthesis for Diffusion Model Personalization

Ethan Smith[1], Rami M. Seid[2]*, Alberto Hojel[3]* ,Paramita Mishra[4], Jianbo Wu[5]

[1]Leonardo AI Research Lab, Sydney, NSW, Australia

[2]Lucid Simulations, Corp

[3]Lucid Simulations, Corp(151)

[4]Precigenetics Inc, California, Merced

[5]SimpleBerry Research Lab United States University of California, Merced

Citation: Smith E, Seid RM, Hojel A, Mishra P, Wu J. Lora Diffusion: Zero-Shot Lora Synthesis fsssor Diffusion Model Personalization. *Int J Cur Res Sci Eng Tech* 2024; 7(4), 111-115. DOI: doi.org/10.30967/IJCRSET/Alberto-Hojel/151

Received: 21 December, 2024; **Accepted:** 24 December, 2024; **Published:** 26 December, 2024

***Corresponding author:** Alberto Hojel, Lucid Simulations, Corp. North, E-mail: alberto@lucidsim.com

ABSTRACT

Low-Rank Adaptation (LoRA) and other parameter-efficient fine-tuning (PEFT) methods provide low-memory, storage-efficient solutions for personalizing textto-image models. However, these methods offer little to no improvement in wallclock training time or the number of steps needed for convergence compared to full model fine-tuning. While PEFT methods assume that shifts in generated distributions (from base to fine-tuned models) can be effectively modeled through weight changes in a low-rank subspace, they fail to leverage knowledge of common use cases, which typically focus on capturing specific styles or identities. Observing that desired outputs often comprise only a small subset of the possible domain covered by LoRA training, we propose reducing the search space by incorporating a prior over regions of interest. We demonstrate that training a hypernetwork model to generate LoRA weights can achieve competitive quality for specific domains while enabling near-instantaneous conditioning on user input, in contrast to traditional training methods that require thousands of steps.

1. Introduction

The emergence of diffusion models[1] has revolutionized image generation, enabling the creation of highly realistic and diverse images, and the integration of text conditioning enables an interface for generating complex scenes guided by user-provided descriptions. However, certain identities or content types remain difficult to generate through text prompts alone, requiring solutions that allow users to specify their desires beyond what can be described in words. This has led to developments in model fine-tuning and other personalization methods[2]. Typically, users fine-tune the model on reference images sharing a consistent element, such as a specific subject or style, enabling the model to generate this content in novel scenarios guided by user prompts.

Traditional fine-tuning can become expensive when serving users at large. It requires significant training time and yields full-size copies of the original model weights, even when learning a single concept. This has spurred development of Parameter-Efficient Fine-Tuning (PEFT) methods[3] that train fewer parameters, dramatically reducing storage costs while maintaining comparable performance for many use cases. Nevertheless, the training process remains computationally expensive, often requiring thousands of steps. While adapter methods (Ye et al., 2023) address this by training additional

conditioning layers once for zero-shot image conditioned generation, they typically sacrifice identity fidelity to reference images and/or limit generalization to new text-described scenarios.

Motivated by the speed of adapter approaches and the quality of PEFT methods, we seek methods that can achieve the best of both worlds. We observe that while PEFT methods can be fine-tuned towards any content, covering an expansive search space of possible solutions (including even random noise or blank images), in practice, users typically fine-tune on a limited subset of this domain such as objects, people, and styles. The domain of user fine-tuning images represents only a slice of the real image manifold, which itself is merely a slice of all possible pixel combinations. Thus, much of this broad fitting capacity goes unused. This observation leads us to explore methods of trading off expressivity for inductive biases that constrain our optimization space towards commonly observed types of content. We accomplish this by establishing priors over our search space, first building a dataset of LoRA adapters trained on our target domain, then training a hypernetwork to synthesize novel, unseen LoRAs. We explore three approaches: First, we propose generating new LoRAs as linear combinations of existing LoRAs, with weight coefficients learned through gradient descent, similar to works discovering learnable merge strategies via evolutionary algorithms[4]. Second, we investigate training a variational autoencoder (VAE)[5] on our LoRA dataset and optimizing a latent vector that, when decoded, produces the optimal fine-tuned LoRA. Finally, we propose a diffusion-based hypernetwork[6].

In our work, we focus on facial identities as a cost-effective domain for experimenting with various strategies.

Our approach leverages conditional diffusion with ArcFace (Deng et al., 2019) embeddings along with a trained VAE.

Our main contributions are:

- We develop training-free methods for sampling new LoRAs, enabling rapid adaptation of text-to-image models
- We implement conditional sampling of LoRAs based on ArcFace embeddings.
- We train a VAE to learn a compact latent representation of our LoRA dataset, facilitating efficient generation and manipulation of new LoRAs

2. Preliminaries And Related Work

2.1. Image Diffusion Models

Diffusion models have become a standard for for image generation. Latent Diffusion models, notably Stable Diffusion[7], demonstrated improved training and inference efficiency by performing the denoising process in the latent space of a Variational Autoencoder[5], followed by decoding back to pixel space.

2.2. Personalization Via Finetuning

Dreambooth[2] proposed a method for fine-tuning Stable Diffusion to capture a person's identity and generate them in novel text-described scenarios. Subsequent approaches have focused on reducing trainable parameters by approximating weight changes through factorized matrices. LoRA, a Parameter-Efficient Fine-Tuning (PEFT)[3] method, optimizes a low-rank parameter difference matrix. For a given matrix W in R^{NxN}, instead of optimizing W directly to obtain converged W^*, matrix ΔW is optimized such that $W^* = W + \Delta W$. By itself, this does not yield any benefits to reducing training costs. However, one can perform a low rank approximation of ΔW by factorizing $\Delta W \approx AB$ where A lies in R^{NxM} and B lies in R^{MxN} and $M \ll N$.

Figure 1: Samples generated from LoRA-adapted Stable Diffusion, where LoRAs are generated by a hypernetwork taking faces as input conditions. Cropped faces show the reference image, and the paired image on the right shows the generated sample.

2.3. Zero-Shot Personalization

IP-Adapter[8] learns an additional set of cross attention weights to enable conditioning on CLIP image embeddings, allowing generations to be guided by a reference image. [9] Subject Diffusion uses dense spatial features which can yield generations with high fidelity to the provided subject, but at the cost of poor flexibility to novel poses and poor prompt following[10]. Arc2Face trains a foundational image diffusion model conditioned on[11] ArcFace embeddings allowing high-fidelity generation of variations of a provided face.

2.4. Hypernetworks

Hypernetworks[6] are a technique involving the training of a network to generate the weights for another neural network. Hypernetworks optimize a weight generator network instead of the child network, thus the child network's work is to propagate weights appropriately.

3. Methodology

Figure 2 illustrates the design of LoRA Diffusion. LoRA Diffusion consists of two phases: Data Collection and Training.

3.1. Problem Statement

We seek to accelerate LoRA fine-tuning methods by discovering a lower dimensional manifold in the LoRA parameter space.

For a given LoRA parameter space Φ in R^N, we aim to uncover a manifold M of dimensionality R^R that resides in Φ where $R \ll N$.

M is defined by a dataset of LoRAs $\{L_1, L_2, ..., L_N\}$ belonging to a shared domain. This could be a specific domain, such as the domain of face images, or broad, such as the domain of real images.

We then define a generative model that takes dataset of LoRAs and an optional condition and samples a new LoRA $p(x|M,c)$.

3.2. Data Collection

Our procedure begins with collecting a dataset of LoRA

adapters trained on the Stable Diffusion model[7]. We utilize the dataset from Weight2Weight[12], which comprises 64k LoRAs, each trained on images of a different celebrity and containing approximately 99k parameters. For detailed information about the Weight2Weight dataset creation, we refer readers to the original paper[12]. We observe that the A and B vector components of LoRA can exhibit varying L2 norms across different trained models. To achieve uniform representations across fine-tuned adapters, we fuse the A and B components by computing their outer product, resulting in a matrix matching the full weight dimensions. **(Figure 2)**

$$W\hat{} = AB, \quad W\hat{} \in \mathrm{R}NxN, \quad A \in \mathrm{R}NxM, \quad B \in \mathrm{R}MxN$$

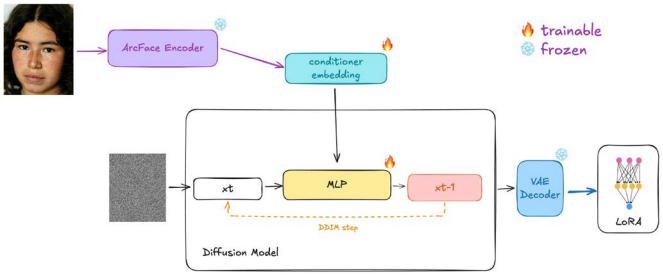

Figure 2: Design of LoRA Diffusion. A frozen VAE encoder is used to encode LoRAs into a latent space of reduced dimensionality. In a training step, gaussian noise is applied to a latent sample, and the MLP is tasked with predicting the denoised LoRA given the ArcFace embedding condition. At inference time, this process begins with a gaussian noise sample and iteratively denoised to produce a latent which is then decoded to a novel LoRA.

This is then followed by the Singular Value Decomposition to obtain U, S, V matrices. The first R components of U and V, where R is the rank of the original trained LoRA, are extracted to become the new parameterization for the LoRA. Finally, each singular vector from each U and V is scaled by the square root of its singular value, thus ensuring equal parameter norm in the resulting A and B LoRA vectors.

$$U,S,V = SVD(W\hat{})$$

$$\sqrt{}$$

$$A\hat{} = U * \quad S\sqrt{}$$

$$B\hat{} = V * S$$

This reparameterization maintains equivalent functionality of the LoRA while standardizing statistics across matrices for modeling purposes. Following this transformation, we flatten the A and B components for each layer into vectors and concatenate them into a single large vector to serve as output targets for the hypernetwork.

$$V = [A_1,B_1,A_2,B_2,...A_N,B_N]$$

We maintain a mapping indicating which feature dimensions correspond to each weight so that generated samples can be unflattened and reassembled into the original LoRA structure. Our LoRA models are trained on 64,000 unique identities derived from the CelebA dataset.

3.3. VAE Training

We trained a variational autoencoder to encode the flattened LoRA vectors into a compressed representation. Initial experiments with a KL divergence weight of 1.0 yielded poor reconstructions. After optimization, we found that a lower KL weight (Beta < 1) improved reconstruction quality while maintaining latent space structure.

The integration of 64k identity LoRAs and CelebA binary features provides the foundation for subsequent training. The final stage involves training a diffusion model on the N-dimensional vector space, resulting in strong conditioning and consistent LoRA generation.

3.4. Adalora

We propose a novel method of conditioning intended to provide more expressivity than methods like AdaNorm[13] which are limited to elementwise scale and shift operations.

Meanwhile, AdaLoRA projects the condition into a transformation matrix to be applied to the hidden states of the network, decomposed into two low rank matrices as per typical LoRA paradigm. Let f(y) represent a function, parameterized as a small neural network, that maps a condition vector in R^m to a matrix A of shape R^{dxr}. Likewise, a similar function g(y) which maps the input to a matrix B of shape R^{rxd}

$$A = f(y), B = g(y), \quad y \in R^m, \quad A \in R^{dxr}, \quad B \in R^{rxd}$$

Following, the hidden states are transformed given the resulting matrices

$$x\hat{} = xAB$$

4. Experiments

This section details a series of experiments conducted to explore novel approaches in latent space manipulation, dimensionality reduction, and generative modeling within the context of LoRA vectors and facial representation learning.

4.1. Latent Space Compression Via Principal Component Analysis (PCA)

We investigate the linear properties of structured latent LoRA vectors by applying PCA to jointly embed these vectors into a covariance matrix. We break this procedure into following steps:

We construct a covariance matrix Σ from $N = 64{,}000$ LoRA vectors $\mathbf{x}_i \in R^d$, where d is the dimensionality of the LoRA vectors.

We perform eigendecomposition on Σ to obtain eigenvectors $\mathbf{v}_j$ and corresponding eigenvalues λ_j.

We select the top $k = 10{,}000$ principal components for dimensionality reduction.

We project LoRA vectors onto the reduced subspace: $\mathbf{z}_i = \mathbf{V}^T\mathbf{x}_i$, where $\mathbf{V} = [\mathbf{v}_1,...,\mathbf{v}_k]$.

The PCA-based compression enables feature disentanglement, facilitating independent manipulation of latent attributes. **(Figure 3)** reveals that the first 10,000 principal components accounted for approximately 95% of the total variance in the LoRA vector space.

However, we discover that the LoRA vectors exhibited non-linear characteristics when attempting to scale the dataset beyond 64,000 samples. This non-linearity results in identity LoRAs deviating from the true human identity subspace, leading to feature interference and reduced fidelity in facial attribute manipulation.

Cumulative Explained Variance Ratio vs. Number of Components

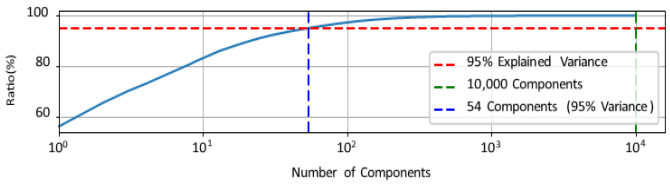

Figure 3: Cumulative explained variance ratio for the top 10,000 principal components of the LoRA latent space, highlighting the diminishing returns of additional components.

4.2. Latent Space Compression Using Variational Autoencoder (VAE)

To address the limitations of PCA, we explore VAE-based latent space compression, which can capture non-linear relationships in the data. We break this procedure into following steps:

- We implement a VAE architecture with:
- Encoder $q_\phi(\mathbf{z}|\mathbf{x})$: Single fully-connected layer mapping input $\mathbf{x}$ to latent mean μ and log-variance $\log\sigma^2$.
- Decoder $p_\theta(\mathbf{x}|\mathbf{z})$: Single fully-connected layer mapping latent $\mathbf{z}$ to reconstructed $\mathbf{x}$.
- We optimize the Evidence Lower Bound (ELBO):

$$\theta,\phi;\mathbf{x}) = E_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) \quad (1)$$

where β is a hyperparameter controlling the KL-divergence weight.

- We utilize the reparameterization trick: $\mathbf{z} = \mu + \sigma \odot \varepsilon$, where $\varepsilon \sim N(0,\mathbf{I})$.
- We experiment with various β values and latent space dimensionalities.

We find that a smaller KL-divergence weight ($\beta < 1$) and a latent space dimensionality of $m = 512$ provided the optimal trade-off between reconstruction fidelity and latent space structure. This configuration results in a more effective latent representation $\mathbf{z} \in R^m$, which we subsequently used in our diffusion model experiments **(Figure 4)**.

The VAE approach demonstrated superior performance in capturing non-linear relationships compared to PCA, allowing for more nuanced facial attribute manipulation and improved interpolation between identities in the latent space.

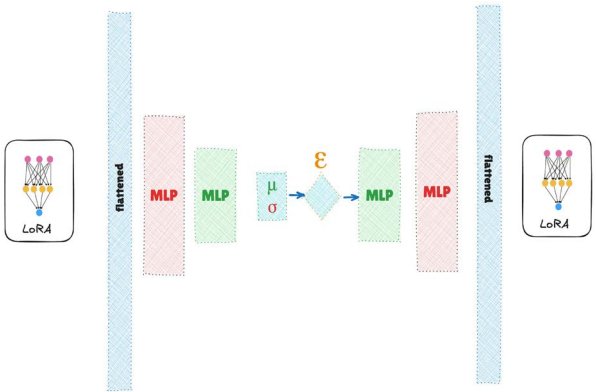

Figure 4: (a) VAE architecture diagram. LoRA weights are flattened into one large vector and fed through sequential MLPs of progressively decreasing dimensions in the encoder, and expanded back to original size in the decoder.

4.3. Diffusion Models with X_0 Prediction On Scaled Loras And Vae Latent

We investigate the efficacy of diffusion models for generating high-quality facial representations using both scaled LoRA vectors and VAE latents. We break this procedure into following steps:

We implement a U-Net-based diffusion model with x_0 prediction:

$$q \qquad q\varepsilon_\theta(\mathbf{x}_t,t) \approx (\mathbf{x}_t - \quad 1-\bar{\beta}_t\mathbf{x}_0)/ \qquad \bar{\beta}_t \qquad (2)$$

where $\mathbf{x}_t$ is the noisy input at timestep t, $\bar{\beta}_t$ is the cumulative noise schedule, and θ are the model parameters.

We apply the model to both scaled LoRAs and VAE latents.

We utilize a cosine noise schedule as proposed by Nichol & Dhariwal (2021).

We analyze the learning dynamics across different layers of the U-Net (Ronneberger et al., 2015) architecture using gradient flow visualization techniques.

Contrary to our initial hypothesis, we observe that LoRA matrices across different layers exhibited high variability, leading to conditioning failures. The model's learning is concentrated in the encoder and decoder layers, neglecting the ResNet[14] and information bottleneck components of the U-Net architecture. This results in suboptimal information flow and poor performance, particularly in preserving identity-related features.

In contrast, VAE latents demonstrates a Gaussian-prior structure, which proved more suitable for the diffusion model's score function. This structure allows for effective interpolation between concepts and aligned well with the diffusion process of reducing high-frequency features before addressing lower-frequency components.

We now present the quantitative results comparing our VAE approach versus using scaled LoRA vectors in **(Figure 5)**. We observe that the VAE method reaches a substantally lower loss value (MSE) on the training set, and substantially higher ArcFace Similarity score on the validation set. We hypothesize this is due to the greater expressivity of the non-linear VAE layers as opposed to linearly combining scaled LoRA vectors.

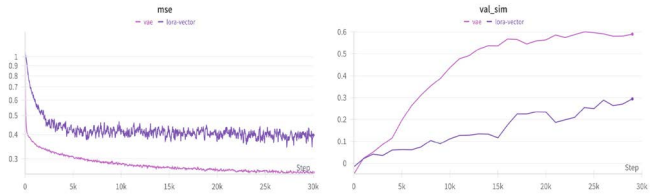

Figure 5: Comparison of diffusion model performance on LoRA vectors vs. VAE latents. (a) Loss curve (b) Validation similarity with arcface embeddings

4.4. Diffusion Models With V-Prediction

Building on our findings from x_0 prediction, we explore v-prediction as an alternative approach, which has shown promise in recent literature for its stability and sample quality. We implement a diffusion model with v-prediction, where α_t and σ_t are time-dependent scaling factors.

$$v_\theta(\mathbf{x}_t,t) \approx \alpha_t\varepsilon - \sigma_t\mathbf{x}_0 \qquad (3)$$

The v-prediction approach demonstrates superior performance compared to x_0 prediction, both in terms of sample quality and training stability. Provides a closer interpretation of flow models in a diffusion setting, learning the score between $\mathbf{x}_t$ and $\mathbf{x}_{t-1}$ using a transformation-like method.

4.5. Adalora: An Alternative To Adanorm For Feature Modulation

In our final experiment, we investigate ADALoRA (Adaptive Low-Rank Adaptation) as an alternative to AdaNorm for conditional feature modulation in our diffusion model. We break this procedure into the following steps:

- We propose a new method, ADALoRA, implemented as follows:

$$\mathbf{h}\text{out} = \mathbf{W}\mathbf{h}\text{in} + \mathbf{B}\mathbf{A}\mathbf{h}\text{in} \qquad (4)$$

where **W** is the original weight matrix, **B** and **A** are low-rank matrices, and $\mathbf{h}_{in}$ and $\mathbf{h}_{out}$ are hidden input and output states. **(Figure 6)**

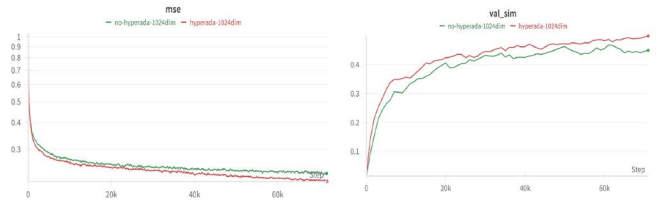

Figure 6: Comparison of ADALoRA vs. AdaNorm. (a) Loss curves. (b) Similiarity scores over training iterations.

- We compare the performance against AdaNorm in terms of conditioning efficacy and generalization.
- We analyze overfitting tendencies using validation loss curves.
- We evaluated sample quality and attribute consistency using arcface similarity score.

ADALoRA demonstrates improved utilization of conditioning information compared to AdaNorm. Allows for more flexible and independent feature modulation across layers, resulting in finer control over generated attributes. We observed a 30% improvement in ARCFace similiarity score for conditional generation (see Figure 6)

The improved performance of ADALoRA can be attributed to its ability to learn more expressive transformations through its low-rank adaptation mechanism, allowing for better fine-tuning of pretrained weights in the context of our facial generation task.

5. Conclusion

We propose a new approach, LoRA Diffusion, which enables zero-shot LoRA synthesis for diffusion model personalization.

Our extensive experimentation with various latent space compression techniques and diffusion model architectures has yielded several key insights for fast and high-fidelity model personalization methods. The combination of VAE-based latent compression, diffusion models, and ADALoRA for feature modulation has shown particular promise for scaling such an approach to larger and more broad domains.

6. Acknowledgments

We sincerely thank the weights 2 weights team for their collaboration on the details of this paper and Amil's team for providing us with the LoRAs.

7. References

1. https://arxiv.org/abs/2006.11239
2. https://arxiv.org/abs/2208.12242.
3. https://github.com/huggingface/peft
4. https://arxiv.org/abs/2403.13187.
5. Diederik P Kingma. Auto-encoding variational bayes, 2013.
6. https://arxiv.org/abs/1609
7. Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-¨ resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022;10684-10695.
8. https://arxiv.org/abs/2308.06721
9. https://arxiv.org/abs/2307
10. https://arxiv.org/abs/2403
11. Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019;4690-4699.
12. Amil Dravid, Yossi Gandelsman, Kuan-Chieh Wang, Rameen Abdal, Gordon Wetzstein, Alexei A Efros, and Kfir Aberman. Interpreting the weight space of customized diffusion models, 2024.
13. https://arxiv.org/abs/2210.06364.
14. Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2016;770-778.