LoRA-Driven Anime Style Generation: A Comparative Study of Lightweight Fine-Tuning Techniques

Zhaokun Wang
4742264
Computational Linguistics

Jiaman Sun
4742476
Scientific Computing

Abstract

We investigate the use of Low-Rank Adaptation (LoRA) for fine-tuning diffusion models on small anime-style datasets. We design a series of experiments to systematically evaluate LoRA's capabilities. First, we benchmark LoRA across multiple styles to assess its generalization under limited data. Second, we conduct hyperparameter ablation by varying rank and learning rates to analyze their influence on performance. Third, we compare LoRA with Textual Inversion and DreamBooth, highlighting the trade-offs between lightweight and full-parameter fine-tuning methods. Finally, we explore style mixing by interpolating LoRA models trained on different styles, examining its compositional ability. Together, these experiments provide a comprehensive view of LoRA's strengths and limitations in low-data fine-tuning scenarios and offer insights into its potential for efficient diffusion model adaptation.

1. Introduction

Diffusion models have achieved state-of-the-art performance in image generation, characterized by high-quality outputs, strong diversity, and the ability to incorporate textual conditions for controllability. Stable Diffusion represents a specific implementation of this framework. Unlike traditional diffusion approaches, it does not perform noise addition or removal directly in pixel space. Instead, it first compresses images into a low-dimensional latent space using a Variational Autoencoder (VAE) before applying diffusion, which results in greater computational efficiency and reduced memory consumption. Stable Diffusion leverages textual prompts to guide image synthesis. However, it exhibits certain limitations: in particular, it demonstrates instability in style generalization and struggles to consistently capture specific artistic domains, such as anime aesthetics.

To address this issue, researchers have proposed various lightweight fine-tuning methods, enabling Stable Diffusion to adapt to specific styles even under limited computational resources and small-scale datasets. Among these approaches, Textual Inversion represents new concepts by learning additional text embeddings, DreamBooth improves stylistic fidelity through broader model parameter fine-tuning, while LoRA (Low-Rank Adaptation) achieves style transfer by introducing low-rank matrix decomposition into selected layers, requiring the training of only a small number of additional parameters. In this project, we primarily investigate LoRA and systematically evaluate its performance on anime-style datasets through a series of experiments, while also conducting comparisons with Textual Inversion and DreamBooth to explore their applicability and potential limitations.

In this project, we focus on investigating the LoRA fine-tuning method and systematically evaluate its applicability in generating anime-style images. Compared to other fine-tuning approaches, LoRA achieves a better balance between training cost and effectiveness, while also possessing the unique advantage of style composability. To comprehensively analyse its performance, we designed four experiments:Exp-I establishes a multi-style benchmark to validate LoRA's generalisation capability across different anime styles; secondly, Exp-II explores the impact of hyperparameters (rank, steps, learning rate) on training outcomes to identify optimal parameter ranges; Exp-III compares LoRA with Textual Inversion and DreamBooth to understand the characteristics and trade-offs of different methods; Finally, Exp-IV demonstrates LoRA's potential for style blending, proving its flexibility in generating hybrid styles. Collectively, these experiments form our systematic research framework for LoRA, providing valuable insights for its application in stylised image generation.

2. Related Work

Diffusion models have recently emerged as the state-of-the-art framework for image generation, offering controllable and diverse synthesis results [2, 4]. While Stable Diffusion provides an efficient latent-space implementation, adapting such large-scale models to specific artis-

tic domains remains challenging. To address this, several parameter-efficient finetuning methods have been proposed. Textual Inversion learns additional embeddings to introduce new styles [1], while DreamBooth finetunes a larger portion of model weights to achieve strong personalization [5].

LoRA (Low-Rank Adaptation) offers a more efficient alternative by inserting low-rank matrices into selected layers [3]. This approach requires training only a small set of parameters, making it suitable for limited datasets and fast adaptation, especially in stylized domains such as anime or illustration. Recent works further explore LoRA's trade-offs in rank, learning rate, and placement, highlighting its potential for flexible style adaptation [6].

Building on these developments, our work systematically evaluates LoRA in the context of anime-style image generation. Through four experiments, we investigate (i) multi-style generalization across different anime aesthetics, (ii) hyperparameter sensitivity, (iii) comparisons with Textual Inversion and DreamBooth, and (iv) style composability via LoRA blending. Together, these studies aim to provide practical insights into LoRA's performance and limitations in stylized diffusion finetuning.

3. Approach

A.Base Model and LoRA Fine-Tuning

We adopt Stable Diffusion v1-5 as our base model. This model is built upon the Latent Diffusion Model (LDM) framework, where images are first compressed into a latent space using a VAE before applying the diffusion process, significantly reducing computational cost. It is particularly suitable for training on anime and illustration styles, and is compatible with various lightweight fine-tuning approaches such as LoRA, DreamBooth, and Textual Inversion.

B. LoRA Fine-Tuning

LoRA fine-tuning applies low-rank decomposition to selected weight matrices $W \in \mathbb{R}^{d\times k}$:

$$W' = W + \Delta W, \quad \Delta W = AB^T \tag{1}$$

where $A \in \mathbb{R}^{d\times r}, B \in \mathbb{R}^{k\times r}$ and $r \ll \min(d, k)$. This formulation drastically reduces the number of trainable parameters compared to directly training ΔW, resulting in faster training and lower memory usage.

In our project, LoRA modules are inserted into the attention layers of the U-Net in Stable Diffusion v1-5. By training only the additional low-rank parameters, we are able to efficiently adapt the model to different anime styles with limited computational resources.

C.Experimental Settings

For our experiments, we curated three distinct anime-inspired styles: Ghibli, Shinkai, and American Comic. Each dataset consists of approximately 100 high-quality images, which serve as the training data for fine-tuning. To ensure fair evaluation across styles, we employed prompts written in a consistent descriptive style across the three datasets.

The key hyperparameter configurations include the rank r, the number of training steps, and the learning rate (lr), which were systematically varied in different experiments. All models were trained using NVIDIA P100 GPU.

D. Compared Methods

We systematically compared LoRA, Textual Inversion (TI), and DreamBooth under the same dataset and prompt settings.

Textual Inversion (TI) encodes style by learning new textual embeddings that can be incorporated into prompts, enabling the base model to synthesise stylised images without modifying its internal parameters.

DreamBooth fine-tunes a substantially larger subset of model parameters with only a few reference images, yielding higher style fidelity but at the cost of increased computational requirements and a higher risk of overfitting on small datasets.

LoRA integrates low-rank adaptation modules into the attention layers of Stable Diffusion's U-Net, updating only a small number of additional parameters. This approach achieves a favourable trade-off between efficiency and effectiveness, while also supporting composability of styles.

By contrasting these three approaches, we aim to highlight their respective strengths and limitations in terms of computational cost, data efficiency, and style generalisation, thereby offering practical insights for anime-style image generation tasks.

E. Evaluation Metrics

To comprehensively evaluate the effectiveness of fine-tuning methods, we employed four widely used metrics in generative image research:

Fréchet Inception Distance (FID): Measures the divergence between generated images and the training set in feature space, reflecting image quality and realism.

Inception Score (IS): Assesses both quality and diversity by examining a classifier's recognition confidence and category distribution of generated samples.

CLIP Score: Evaluates semantic alignment between generated images and text prompts, ensuring outputs follow the intended style or content.

LPIPS (Learned Perceptual Image Patch Similarity): Quantifies perceptual diversity by measuring differences between samples within the same style.

Across our four experiments, these metrics were applied as follows:

Exp-I: Used FID and CLIP to validate LoRA's generalisation across multiple styles.

Exp-II: Analysed the effect of hyperparameters (rank, steps, lr) on performance, identifying optimal ranges.

Exp-III: Compared LoRA, Textual Inversion, and DreamBooth, highlighting differences in efficiency, data usage, and style fidelity.

Exp-IV: Relied on CLIP and LPIPS to assess style fusion effectiveness and intra-style diversity.

4. Experiments

A.Exp-I: Multi-Style LoRA Benchmark

We curated three anime-inspired datasets (Ghibli, Shinkai, American Comic), each containing 100 high-quality images covering both character and background elements. All images were resized to 512 × 512 resolution, following the default Stable Diffusion configuration.

For LoRA fine-tuning, we adopted rank = 16, training steps = 1000, and learning rates of 1e-4 for both U-Net and the text encoder. All experiments were conducted on a single NVIDIA P100 GPU.

We used the original captions associated with each training image as prompts. The LoRA models fine-tuned on each dataset were asked to re-generate the corresponding images. The generated results were then compared with the original images to evaluate reconstruction quality using FID, KID, CLIP, and LPIPS. For comparison, the pretrained Stable Diffusion model without fine-tuning was also evaluated under the same prompts.

The objective of Experiment I is to assess LoRA's ability to faithfully reconstruct stylistic features within each dataset, under the limited-data setting, across three distinct anime-inspired styles.

B.Exp-II: Hyper-Parameter Ablation (Ghibli)

The purpose of Exp-II is to systematically analyse the influence of key hyperparameters on the performance of LoRA fine-tuning. Specifically, we varied the rank (r) of the low-rank matrices, the number of training steps (s), and the learning rates (lr) applied to the U-Net and text encoder.

Rank (r): We experimented with values of r = 8, 16, 32.

Training steps (s): Each model was trained for s = 3000 iterations.

Learning rates (lr): We tested two configurations:

$$\text{lr}_{\text{unet}} = 1 \times 10^{-4}, \quad \text{lr}_{\text{text}} = 1 \times 10^{-4}$$

$$\text{lr}_{\text{unet}} = 5 \times 10^{-4}, \quad \text{lr}_{\text{text}} = 5 \times 10^{-4}$$

All other experimental conditions remained consistent with Experiment I, including the dataset (100 images per style) and uniform text prompts. Through this controlled ablation study, Exp-II aims to investigate the impact of rank, training steps, and learning rate on model convergence speed, image quality, and style fidelity, thereby providing a reference for hyperparameter selection in LoRA fine-tuning.

C.Exp-III: LoRA vs. TI vs. DreamBooth

We systematically compared three popular Stable Diffusion fine-tuning methods: LoRA, Text Inversion (TI), and DreamBooth. To ensure a fair comparison, we employed a Ghibli-style dataset comprising 100 curated images and applied an identical set of prompts across all methods.

The number of training steps was fixed to 1000 for all experiments. For LoRA, the rank was set to r = 32, while TI and DreamBooth do not involve rank hyperparameters.

Since the optimization dynamics of the three approaches differ substantially, we employed distinct learning rate configurations:

LoRA: learning rate for both U-Net and text encoder set to 1e-4.

Textual Inversion(TI): text encoder learning rate set to 5e-3, reflecting the small number of trainable parameters.

Dream Booth: U-Net learning rate set to 1e-4 and text encoder learning rate set to 5e-6, in order to avoid overfitting when fine-tuning a large subset of parameters.

D.Exp-IV: Style Mixing

This experiment investigates the compositional capability of LoRA in combining distinct styles. We selected LoRA models fine-tuned on different anime styles from Exp-I (e.g., *Ghibli* and *Shinkai*) and interpolated their weights during inference. Specifically, for the corresponding parameters of the U-Net and text encoder, θ_A and θ_B, we computed mixed weights as

$$\begin{aligned} \theta_{\text{mix}} &= \alpha\theta_A + (1-\alpha)\theta_B, \\ \alpha &\in \{1.0, 0.75, 0.5, 0.25, 0.0\}. \end{aligned} \tag{2}$$

For each run, a single text prompt was fixed, and images were generated across different values of α, resulting in a sequence that visualises the gradual transition from one style to another under the same semantic content. To further validate the generalisation of this approach, we repeated the experiment with multiple prompts while keeping the hyperparameters constant (rank = 16, steps = 1000, learning rate =1e-4).

5. Results

A.Exp-I: Multi-Style LoRA Benchmark

For each style dataset, images were regenerated from their original captions using LoRA and compared with the originals. We report FID, CLIP, IS, and LPIPS, with the pretrained Stable Diffusion model as baseline.

Style	Method	FID	CLIP	IS	LPIPS
Ghibli	Baseline	235.30	33.13	5.92	0.75
	LoRA	188.23	32.93	4.36	0.65
	Δ	-47.07	-0.21	-1.56	-0.10
Shinkai	Baseline	239.29	32.73	5.26	0.75
	LoRA	210.47	31.66	4.55	0.65
	Δ	-28.82	-1.06	-0.71	-0.09
American Comic	Baseline	221.28	33.79	4.76	0.75
	LoRA	201.24	33.39	4.00	0.70
	Δ	-20.03	-0.40	-0.76	-0.06

Table 1. Experiment I results across three style datasets. LoRA is compared with the pretrained Stable Diffusion baseline. Δ denotes LoRA − Baseline.

As shown in Table 1, compared with the pretrained Stable Diffusion baseline, the LoRA models consistently achieve lower FID scores (-47.07 for Ghibli, -28.82 for Shinkai, and -20.03 for American Comic), indicating that the fine-tuned models generate images closer to the original style distributions.LoRA can capture and reproduce style-specific features under limited-data conditions.

However, CLIP scores decrease slightly across all three datasets, suggesting weaker alignment between generated images and textual descriptions. IS and LPIPS diversity also drop, reflecting reduced diversity in the outputs.This limitation may be due to the small dataset size and insufficient training steps.

From Figure 1, it can be observed that LoRA captures certain stylistic characteristics, such as color tones and overall atmosphere. However, the generated objects and character details are insufficient, which is consistent with the quantitative results reported in Table 1.This suggests that LoRA may exhibit certain limitations when trained on small datasets and under limited training steps.

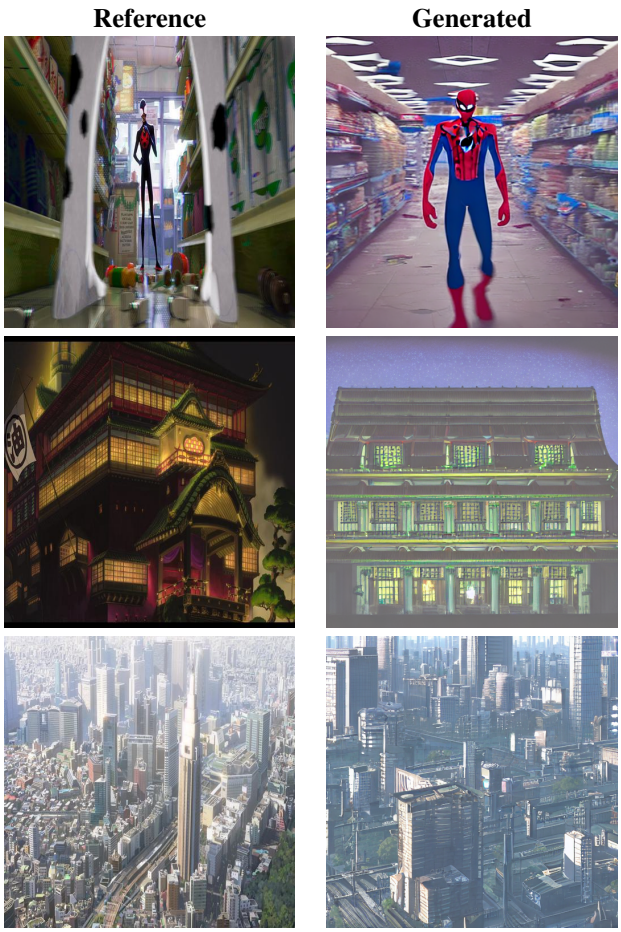

Figure 1. Experiment I : Comparison between reference images (left) and generated results (right) for three different styles.

B.Exp-II: Hyper-Parameter Ablation (Ghibli)

To further investigate the effect of hyperparameters on LoRA fine-tuning, we designed Experiment II by varying the rank (r) and the learning rate (lr). In contrast to Experiment I, which used only $s = 1000$ steps, we increased the number of training steps to $s = 3000$ in order to provide a more comprehensive evaluation of the influence of different hyperparameter settings. The results are reported in Table 2.

r	lr	FID	CLIP	IS	LPIPS
8	1×10^{-4}	179.05	32.56	3.95	0.594
8	5×10^{-4}	180.15	31.57	4.02	0.641
16	1×10^{-4}	185.35	32.50	4.11	0.603
16	5×10^{-4}	175.05	31.70	3.96	0.633
32	1×10^{-4}	182.35	32.47	4.26	0.600
32	5×10^{-4}	177.33	31.60	4.34	0.631

Table 2. Experiment II: Hyper-Parameter Ablation on the Ghibli dataset. We vary rank (r) and learning rate (lr) while fixing steps = 3000.

As shown in Table 2, increasing the number of training steps from $s = 1000$ (Experiment I) to $s = 3000$ on the Ghibli dataset leads to a significant reduction in FID (down to 185.35 at $r = 16, lr = 1 \times 10^{-4}$). This confirms that sufficient training iterations are crucial for LoRA fine-tuning under limited data conditions.

In addition, we observe a clear trade-off with respect to the rank (r). When r is too small (e.g., $r = 8$), the model lacks expressive capacity and fails to capture sufficient stylistic features. Conversely, when r is large (e.g., $r = 32$), the model capacity increases but the performance does not consistently improve, and in some cases even degrades. This may suggest a tendency toward overfitting under small-scale data, where $r = 16$ provides the best balance between representation power and generalization ability.

Looking across evaluation metrics, FID improves substantially with longer training and appropriate learning rates, while CLIP alignment slightly decreases. This indicates that generated images become closer to the training distribution but are somewhat less aligned with textual

prompts. Meanwhile, IS and LPIPS remain relatively stable, implying that LoRA's sensitivity primarily lies in image quality and style consistency rather than diversity.

Regarding the learning rate, a higher lr (5×10^{-4}) accelerates convergence and yields lower FID values in some cases, but is also associated with reduced CLIP scores and potential instability. By contrast, a lower lr (1×10^{-4}) provides smoother optimization trajectories, though requiring more iterations to reach comparable performance. This observation highlights the trade-off between convergence speed and training stability.

Overall, the results of Experiment II demonstrate that LoRA's performance on small datasets is highly sensitive to the choice of rank, learning rate, and training steps. Careful hyperparameter tuning not only improves generation quality under data-scarce conditions, but also helps to mitigate risks of overfitting and semantic drift, offering practical guidance for applying LoRA to larger-scale or more complex domains.

C.Exp-III: LoRA vs. TI vs. DreamBooth

To further compare different fine-tuning strategies under limited-data conditions, we evaluate LoRA, Textual Inversion (TI), and DreamBooth in Experiment III. Table 3 reports the results on the Ghibli dataset.

Method	FID	CLIP	IS	LPIPS
LoRA	180.49	33.02	4.34	0.667
Textual Inversion (TI)	200.95	30.26	5.17	0.777
DreamBooth	224.43	28.30	5.31	0.691

Table 3. Experiment III: Comparison of LoRA, Textual Inversion (TI), and DreamBooth on the Ghibli dataset (steps = 1000).

As shown in Table 3, the three methods exhibit distinct behaviors across different evaluation metrics. LoRA achieves the best overall performance, with the lowest FID (180.49) and the highest CLIP score (33.02). This indicates that LoRA generates images that are both closer to the target distribution and more semantically aligned with the text prompts. These results highlight LoRA's ability to maintain a good balance between image quality and textual consistency under limited-data conditions.

Textual Inversion (TI) achieves the highest IS (5.17) and LPIPS (0.777) scores, reflecting greater diversity and variation among generated samples. However, this diversity comes at the expense of fidelity and semantic alignment: TI records worse FID and CLIP compared to LoRA, suggesting that while TI enhances variability, it struggles to consistently preserve the intended style or text-image alignment.

DreamBooth demonstrates yet another trade-off. It achieves the highest IS (5.31) and relatively high LPIPS (0.691), showing strong capability in producing sharp, distinguishable outputs and retaining distinctive features from the training data. This reflects DreamBooth's tendency toward personalization and memorization. Nevertheless, DreamBooth also records the weakest FID (224.43) and the lowest CLIP (28.30), indicating that its outputs deviate significantly from the target distribution and fail to generalize well to new text prompts.

Overall, LoRA provides the most balanced results by combining low FID with strong semantic alignment, TI prioritizes diversity at the cost of fidelity, and DreamBooth emphasizes personalization but suffers from overfitting under small datasets. These findings underscore the importance of selecting finetuning strategies based on task-specific requirements, whether prioritizing quality, diversity, or personalization.

D.Exp-IV: Style Mixing

To further evaluate the compositional ability of LoRA, we conducted a style mixing experiment by interpolating models fine-tuned on different anime styles. Specifically, we blended the weights of the Ghibli and Shinkai models with varying coefficients. Figure 2 illustrates the visual results, where the generated images demonstrate a smooth transition of styles as the mixing parameter α decreases from 1.0 (pure Ghibli) to 0.0 (pure Shinkai).

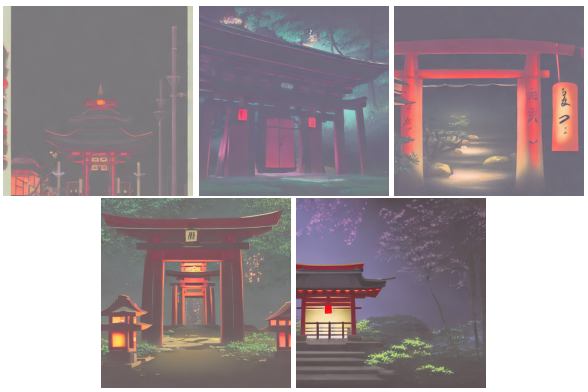

Figure 2. Style mixing results between Ghibli (=1.0) and Shinkai (=0.0). The sequence shows a smooth transition of styles as decreases from 1.0 to 0.0

From Figure 2, we observe that varying the mixing coefficient α from 1.0 (pure Ghibli) to 0.0 (pure Shinkai) produces images that reflect a gradual stylistic shift. At intermediate values (e.g., $\alpha = 0.5$), the generated samples display elements that appear to combine aspects of both styles, such as transitions in color tone and overall atmosphere.

However,the interpolation is not fully smooth. Some generated images show blurred textures, missing edges, or inconsistencies in semantic details, particularly around mid-range α values. This suggests that while LoRA can capture coarse-level stylistic features from both domains, it struggles to maintain fine-grained detail during the blending process. One possible explanation is the limited training data,

which may restrict the model's ability to learn highly detailed and transferable style representations. Alternatively, the interpolation artifacts could also reflect intrinsic limitations of LoRA's parameterization when applied to style mixing.

Overall, Experiment IV indicates that LoRA can achieve a degree of controllable style transition, but the interpolation process still suffers from artifacts and loss of fine details.

6. Conclusion

In this work, we investigated LoRA as a parameter-efficient fine-tuning approach for style adaptation under small-data constraints. Our experiments show that LoRA can capture stylistic elements such as color tone and overall atmosphere, while requiring far fewer parameters than full fine-tuning. These results highlight LoRA's potential as a lightweight method for transferring large diffusion models into artistic domains.

At the same time, our study exposes several limitations. The generated images often lack fine-grained detail, with simplified object structures and incomplete textures. Semantic consistency is not fully preserved: while style is transferred, some outputs deviate from the intended content described by the prompt. Moreover, many samples appear desaturated or grayish, and the improvement in FID is not uniformly strong across settings. Style mixing experiments confirm LoRA's flexibility, but also show intermediate artifacts and inconsistencies during interpolation.

Looking forward, these limitations point to opportunities for refinement. More diverse and larger-scale training data, as well as adaptive scheduling of hyper-parameters, may improve both semantic alignment and detail fidelity. Integrating LoRA with complementary techniques such as Textual Inversion or DreamBooth could further enhance diversity while retaining efficiency. Overall, LoRA represents a promising compromise between efficiency and quality, but further methodological advances are required to achieve high-fidelity, semantically consistent, and detail-rich style generation in practice.

Beyond theoretical insights, this work provides useful directions for applying parameter-efficient finetuning in practice, ranging from anime-style generation and artistic rendering to fast prototyping under restricted computational budgets.

References

[1] Rinon Gal, Or Patashnik, Haggai Maron, Amit Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *arXiv preprint arXiv:2208.01618*, 2022. 2

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 1

[3] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2021. 2

[4] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[5] Nataniel Ruiz, Yuqing Yang, Varun Jampani, Deva Ramanan, Hannaneh Hajishirzi, Abhinav Gupta, and Ting Liu. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *arXiv preprint arXiv:2208.12242*, 2022. 2

[6] Yiming Zhang, Xin Li, and et al. Adaptive lora for efficient fine-tuning of diffusion models. *arXiv preprint arXiv:2305.08672*, 2023. 2