Timestep and Vector Property–Guided LoRA Fusion for training-free style content adaptation

Abstract

The advent of diffusion-based generative models has transformed image synthesis, enabling high-quality and diverse outputs across a wide range of creative domains. As these models become more accessible, personalization has emerged as a key area of interest, particularly for generating images in a specific artistic style or maintaining a consistent character identity across multiple generations with custom-level gpu. Achieving this requires the model to not only adapt to a target style but also preserve content features that define the character's unique appearance. Low-Rank Adaptation (LoRA), a popular parameter-efficient fine-tuning (PEFT) technique, has proven effective for such tasks by enabling targeted adaptation without training entire model parameter. However, combining style and content lora adapter can lead to undesirable effects such as content degradation, structural distortion, and inconsistent stylization To address this, we propose a novel fusion framework that integrates two independently trained LoRA modules—one specializing in style and the other in content and dynamically combines them using singular value decomposition (SVD)-based importance estimation along with a timestep-dependent weight function. This fusion technique is performed training-free, allowing the method to retain the strengths of each module while preventing degradation. Empirical results demonstrate that this approach produces more coherent and visually pleasing outputs, achieving a superior balance between stylistic fidelity and character consistency compared to other baseline approaches.

1. Introduction

Diffusion generative models have rapidly become the dominant approach for high-quality image synthesis and diverse image generation across a wide range of domains. Adapting a pre-trained diffusion model to personalized customization requires model to capture a specific style or maintain a particular character, has attracted increasing attention [4, 8, 13]. However, full fine-tuning of large diffusion models is computationally expensive and memory-intensive, making it impractical for deployment on custom or resource-constrained GPUs. As an alternative, Parameter-efficient fine-tuning (PEFT) techniques such as Low-Rank Adaptation (LoRA) have emerged, enabling the integration of new concepts with minimal additional parameters by decomposing weight updates into low-rank matrices. Specifically, the weight matrix is updated as

$$\mathbf{W}_0 + \mathbf{B}\mathbf{A}, \text{ where } \mathbf{B} \in \mathbb{R}^{d\times r},\ \mathbf{A} \in \mathbb{R}^{r\times k}, \tag{1}$$

where $\mathbf{B}$ and $\mathbf{A}$ are low-rank matrices that approximate the weight update. However, When we combine multiple LoRA adapters, the interaction between their weight adaptations frequently lead to content distortion in the generated outputs. Additionally, mismatched style representations cause to downgrade overall visual fidelity. Previous studies suggests various approach to handle this problem [3, 14, 20]. However, they still suffer from low style transferability and failure to preserve content. To this end, we propose fusion technique which considers property of two LoRA adapters: one dedicated to capturing style, and the other to modeling the content representation. Our fusion process is guided by singular value decomposition (SVD)-based importance estimation and timestep-dependant weight function. Rather than merging these capabilities with training, we combine them during inference, allowing each adapter to specialize while adjusting its contribution according to the depth of the model. This approach enables us to dynamically balance the relative strengths of style and content contributions while preserving the quality of the generated image. Our contributions are summarized as follows:

- **Singular Value Decomposition (SVD) Fusion.** In vanilla fusion, multiple LoRA adapters are simply combined using a linear combination:

$$\Delta W = \sum_{i=1}^{k} \lambda_i \Delta W_i, \tag{2}$$

  where λ_i are user-defined or heuristically chosen factors. Instead of using fixed λ_i,, we propose a more sophisticated fusion strategy that leverages Singular Value Decomposition (SVD) to dynamically determine the hyperparameter λ . Our method is based on the idea that the

fusion of two LoRA adapters should be conditional on their discrepancy. To quantify this discrepancy, we utilize SVD to calculate the difference between two LoRA adapters. We define a threshold, and if the calculated difference is below the threshold δ, we proceed with the fusion by lambda which is proportional to the calculated difference. This approach ensures that important style or content information encoded in different adapters is retained in the fusion.

- **Timestep-dependant weighting** Previous studies [3, 12, 15, 19] have shown that only certain parts of a transformer block are responsible for generating specific image features, and that the regions involved differ between shallow and deep layers of the base model. In particular, content LoRA tends to be more effective when applied to shallow layers, whereas style LoRA achieves better results when applied to deeper layers. Building on these insights, we explore a timestep-based adjustment strategy that selectively retains only the important information for image generation. Specifically, We employed a decreasing exponential function to enhance the strength of the content LoRA in the initial layers, and a logistic function to increase the influence of the style LoRA in the deeper layers.

2. related work

Diffusion model Diffusion models [6, 17] emerged as a powerful approach for image generation tasks. The denoising step encourages the network to generate high-fidelity sample by regressing the added Gaussian noise at each time-step. Additionally, flow matching (also known as rectified-flow diffusion) [2, 10] replaces iterative denoising by directly learning the continuous-time velocity field that transports noise to data and vice versa, yielding straight trajectories and achieving up to fewer sampling steps while preserving sample quality. The advancement of the Diffusion Transformer (DiT) [2, 9, 21], which replaces the traditional U-Net backbone with a transformer architecture operating on latent or image-patch tokens, brought enhanced scalability and state-of-the-art results. Moreover, visual autoregressive models [18], VAR family, have attracted significant attention. VAR generates image via next token prediction over multi-resolution token maps, enabling GPT-style transformers to generate high quality samples.

LoRA in diffusion model. In the field of image generation, achieving both image stylization and content preservation has gained huge interests, with various approaches emerging, such as IP-Adapter [11], DreamBooth [13], customization of diffusion model [8], and textual inversion [4]. One of the most powerful and efficient approaches is LoRA (Low-Rank Adaptation), which has proven highly effective and memory-efficient [1, 7]. Additionally, Researchers are exploring numerous techniques for combining multiple LoRA models. These methods include balancing different LoRAs by learning coefficient parameters [14] and utilizing the top-k elements of a matrix to determine which layers of the diffusion model to modify [12]. Sophisticated methods for concept injection and composition have been developed to generate subjects with intended characteristics [20]. other multi-LoRA composition frameworks employ [23]. other efforts focus on orthogonality and timestep with different LoRA modules [15, 22]. The versatility of these techniques extends beyond static images, finding applications in video diffusion models for tasks like consistent editing and camera movement [5, 22]

3. Methodology

In this section, we introduce our framework, which leverages property of content and lora adapter based on two findings: 1) Naive LoRA fusion with simple weighted summation often produce suboptimal results. 2) The impact of LoRA adapters is layer-dependent, motivating the need for depth-aware fusion strategies.

3.1. Task Formulation

The generation process begins by initializing a latent representation, z_T. Concurrently, a text prompt, p, is processed by a text encoder to produce a conditional text embedding, c. These serve as the initial inputs to the **diffusion transformer block** $\mathcal{D}$, which iteratively denoises the latent representation over a sequence of timesteps, $t \in \{1, 2, \ldots, T\}$. At each timestep, the core operation of the diffusion block can be expressed as:

$$z_{t-1} = \mathcal{D}(z_t, t, c), \tag{3}$$

where z_t denotes the latent representation at timestep t, and z_{t-1} is the denoised latent from the previous timestep. Once the denoising process is terminated, the final latent representation z_0 is passed through a variational autoencoder (VAE) decoder to obtain the final image in pixel space.

3.2. Analysis of LoRA fusion

Our investigation into weight fusion began by exploring simple linear combinations of LoRA adapters. We observed that naive LoRA fusion, such as applying a weighted summation controlled by a scalar hyperparameter (λ) in Equation 2, often produces suboptimal results, as illustrated in Figure 2. We found that indiscriminately merging the entire weights of adapters through simple linear fusion often leads to significant image distortions and a lack of structural coherence. This led us to the finding that the effective application of a LoRA should account for weight mismatches and the inherently timestep-dependent nature of the diffusion process. Different layers are responsible for generating

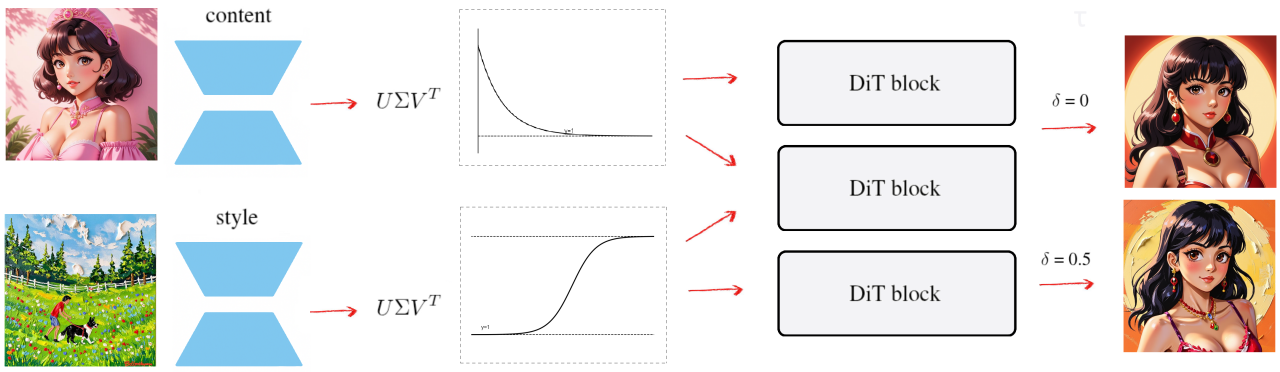

Figure 1. Overview of framework. Content and style LoRA adapters are first decomposed via singular value decomposition (SVD) to capture dominant directions in their weight matrices. The resulting singular vectors inform adapter selection and weighting at each DiT block. By modulating the contribution of content and style adapters across timesteps, the framework enables controlled fusion, producing outputs that balance structural fidelity and stylistic transformation.

various image features at different levels of detail. To validate this and move beyond the limitations of simple summation, we conducted an experiment 4 to analyze the layer-specific behavior of LoRA adapters and determine if a more sophisticated, depth-aware fusion strategy could yield superior results.

3.3. Content and style adapter fusion strategy

To account for the influence of adapters across layers, scaling mismatches between adapters, more sophisticated fusion strategies are required to preserve model performance and avoid degradation. To fully utilize model's capability and adapter, we analyze style and subject lora adapter's weight independently. thereby yield pertinent combination of each lora weight and base model weight. As illustrated in figure , we observe significant norm disparities between style and content lora. To tackle the challenge, we first apply normalization to ensure their values are on a comparable scale Let $\mathbf{S}$ be the style matrix and $\mathbf{C}$ be the content matrix. The Frobenius norms are computed as:

$$\|\mathbf{S}\|_F = \left(\sum_{i,j} S_{ij}^2\right)^{1/2}, \quad \|\mathbf{C}\|_F = \left(\sum_{i,j} C_{ij}^2\right)^{1/2}$$

Define the scaling factor α as the average of the two norms:

$$\alpha = \frac{\|\mathbf{S}\|_F + \|\mathbf{C}\|_F}{2}$$

Normalize both matrices so they have the same norm:

$$\mathbf{S}_{\text{norm}} = \frac{\alpha}{\|\mathbf{S}\|_F + \epsilon} \cdot \mathbf{S}, \quad \mathbf{C}_{\text{norm}} = \frac{\alpha}{\|\mathbf{C}\|_F + \epsilon} \cdot \mathbf{C}$$

Cast both normalized matrices to 32-bit floating point format.

Singular Value Decomposition

After normalization, to determine whether the content-oriented or style-oriented LoRA adapter should be applied at each fusion step, we perform a singular value decomposition (SVD) on the respective weight matrices. From the resulting singular vectors, we compute the element-wise sum for each adapter, serving as a coarse measure of directional bias in the parameter space. Intuitively, a larger absolute sum indicates a stronger and more coherent transformation along a dominant direction, suggesting a higher potential impact on the output representation. At runtime, we first compare the sums from the content and style LoRA adapters. If the difference between them exceeds a predefined threshold, we select the adapter with the larger sum for integration. Otherwise, when the difference is below the threshold, we combine the two adapters using a weighted sum controlled by a hyperparameter λ, which is proportinal to calculated difference $\mathbf{d}$. The standard SVD decomposes matrix A into three matrices:

$$A = U\Sigma V^T$$

Where:

- $U \in \mathbb{R}^{m\times m}$ is an orthogonal matrix (left singular vectors),
- $\Sigma \in \mathbb{R}^{m\times n}$ is a diagonal matrix with singular values σ_i,
- $V \in \mathbb{R}^{n\times n}$ is an orthogonal matrix (right singular vectors).

This provides a complete and robust understanding of the transformations performed by the LoRA adapter, which is a powerful foundation for building more nuanced and effective adaptive strategies

Figure 2. Qualitative comparison of our approach for style-content blending. The first row shows the content and style images to be combined, followed by results from standard LoRA fusion, content-weighted fusion, style-weighted fusion (1.3×), K-LoRA, and our proposed method with thresholds 0 and 0.5. Our method better preserves content while accurately transferring style.

3.4. timestep dependent weight function

Based on analysis in section 4.2, We further employ a timestep-dependent weight function to dynamically leverage the influence of each lora adapter as well. this approach prioritizes the content matrix in the initial layers, where structural and semantic information is established, and gradually shifts the emphasis toward the style matrix in the later layers, where fine-grained textures and stylistic attributes are refined [15] In our proposed framework, we leverage a layer-wise fusion strategy for LoRA weights that is carefully designed to align with the roles of shallow and deep layers in the diffusion process.

Our content lora adapter weight function as follows: timestep parameter

$$S(t) = (\gamma - 1) \cdot \exp\left(-100 \cdot \frac{t}{t_{\text{all}}}\right) + 1$$

where t denotes the current timestep and t_{all} is the total number of timesteps. The parameter γ controls the initial amplitude of the weight, ensures that at $t = 0$ the weight starts near its maximum value. The exponential term $\exp\left(-100 \cdot \frac{t}{t_{\text{all}}}\right)$ introduces a steep decay, rapidly reducing the influence of the content adapter as timesteps progress which converges to 1.

For the style LoRA adapter, we employ a logistic weight function bounded between 1 and 2 to gradually increase its

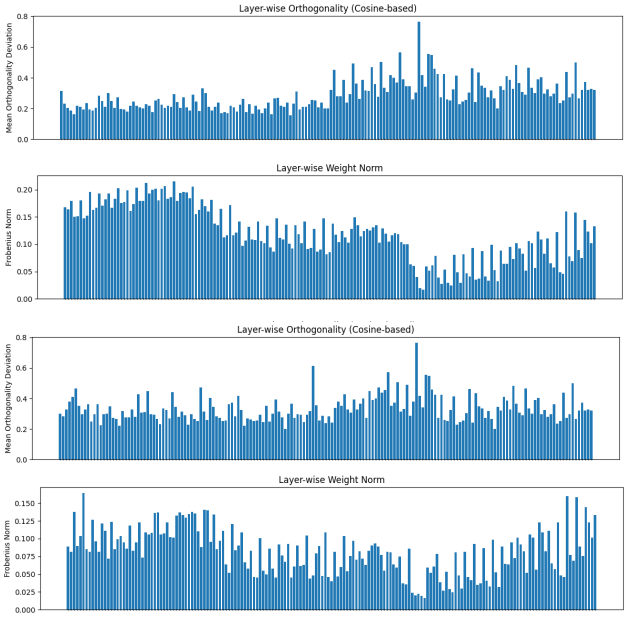

Figure 3. This figure presents a comparative analysis of the layer-wise orthogonality and Frobenius norm of LoRA weights for both content and style models.

influence during generation:

$$S_{\text{style}}(t) = 1 + \frac{1}{1 + \exp\left(-k \cdot \left(\frac{t}{t_{\text{all}}} - \tau\right)\right)},$$

where t is the current timestep, t_{all} is the total number of timesteps, k controls the steepness of the transition, and τ determines the midpoint of the growth. As timesteps progress, the logistic curve increases, approaching 2 toward the end of generation. This design complements the strength of style LoRA adapter in later layers, enabling a smooth and controlled handover from structure to style.
These weight functions are incorporated into the comparison described in Section 3.3, enhancing the likelihood of selecting the content LoRA in the initial layers while progressively increasing the influence of the style LoRA in later stages.

4. Experiment

In this section, we illustrates main results of our experiment to demonstrate the effectiveness of our approach. We employ Flux1-dev as the base model, while using various lora adapter from huggingface
Baseline. We evaluate the performance of our approach against existing methods:(1) **K-LoRA** (2) Naive LoRA fusion (3) content-augmented fusion (4) style-augmented fusion

Model	style similarity	character consistency
base fusion	-	0.4810
style leverage	0.6510	0.3188
content leverage	0.5531	0.7312
K-LoRA	0.6480	0.2803
Ours	0.6412	0.4349

Table 1. Comparison of different fusion strategies in terms of style similarity and character consistency. Style similarity is measured using DINOv2 embeddings, while character consistency is evaluated using CLIP embeddings. Our method achieves a balanced improvement across both metrics compared to existing approaches.

4.1. Quantitative Results.

For quantitative evaluation on generated image, we use three metrics: 1) style similarity score[16], content consistency score, image quality score measured by CLIP and dino-v2. As shown in Table 1, our method yield desirable score compared to others across all metrics. In our evaluation of various generative models, we utilized CLIP and DINOv2 to quantify performance on two critical, often competing, metrics: style similarity and character consistency. As evidenced in the results, a clear trade-off is observed among the baselines. Models optimized for style, such as the style leverage (0.6510), achieved high style similarity scores but suffered from poor character consistency (0.3188 and 0.2803, respectively). Conversely, the content leverage model demonstrated superior character consistency (0.7312) at the cost of its stylistic fidelity. Our proposed model, however, effectively mitigates this trade-off. It achieves a high style similarity score of 0.6412, on par with the specialized style-focused models, wsimultaneously yielding a substantially improved character consistency score of 0.4349. This balanced performance underscores our model's capacity to generate images that are not only stylistically faithful but also robustly preserve the identity of the core character, presenting a more holistic and practical solution for style-conditioned image generation tasks.

4.2. analysis

We first conducted experiments by varying timesteps and progressively increasing the influence of the content LoRA in the initial layers to validate the timestep-dependent effects, as shown in the top row of Figure 4. We considered three scenarios in this study: 1) Baseline, simple weighted summation of LoRA adapters without any layer-specific adjustments. 2) Applying both SVD-based fusion and timestep-dependent weighting, with the threshold adjusted to prioritize content preservation. 3) The fully-equipped setup with a different threshold to emphasize style trans-

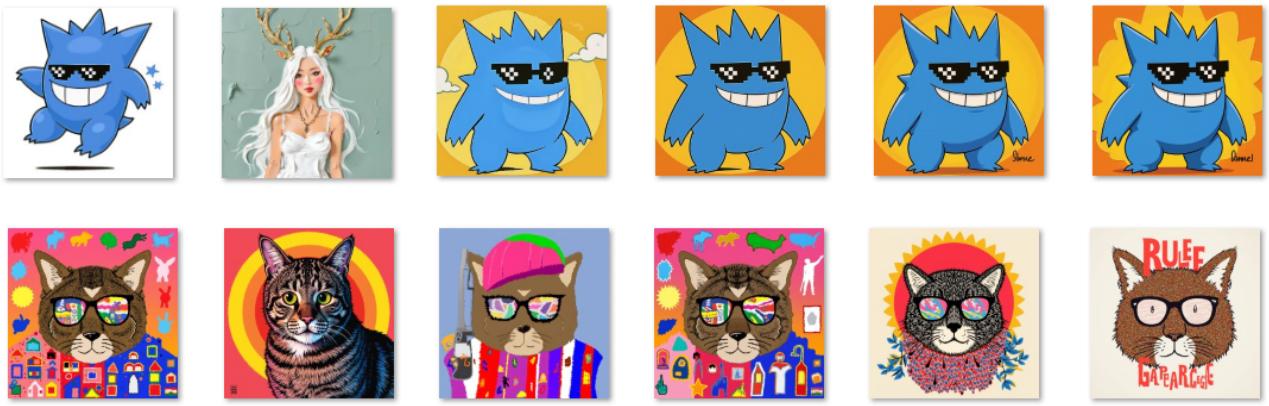

Figure 4. Top row: Visual representation of image shifts in a generated output by increasing the adaptation of a content LoRA in the initial layer. Bottom Row: A comparative study of different LoRA-based generation methods using a fixed seed. The images, from left to right, illustrate outputs from: (a) LoRA Fusion, (b) k-LoRA, (c) Content 1.3x LoRA Fusion, (d) Style 1.3x LoRA Fusion, and (e) our proposed method. These outputs highlight the varying effects of each technique on the final generated image.

fer. The qualitative results are presented in Figure 2. In the baseline, one concept often dominated the other, leading to content distortion. In contrast, the threshold-adjusted fully-equipped settings successfully preserved content while applying the desired style, demonstrating effective trade-offs between content fidelity and style transfer depending on the threshold. This can be clearly observed in the bottom row of Figure4. Using the same seed, the bottom row compares different LoRA-based generation methods, revealing that naive LoRA fusion via simple weighted summation repeatedly causes content distortion. In contrast, our method demonstrates the ability to preserve content while enhancing style transferability.

4.3. conclusion

In this work, we present a manipulatable LoRA fusion framework without additional training by introducing a depth-aware, timestep-dependent strategy that adaptively balances content and style contributions throughout the diffusion process. By normalizing LoRA weights, analyzing their structural differences via singular value decomposition, and applying dynamic weighting functions, we ensure that content LoRA dominates in early timesteps to preserve structural and semantic integrity, while style LoRA gradually gains influence in later timesteps to refine texture and artistic details. This adaptation enables more coherent, visually appealing, and semantically faithful image generation compared to conventional fusion approaches and provide a robust and versatile solution for high-quality controllable generation.

References

[1] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2

[2] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. 2

[3] Yarden Frenkel, Yael Vinker, Ariel Shamir, and Daniel Cohen-Or. Implicit style-content separation using b-lora, 2024. 1, 2

[4] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 2

[5] Chenjian Gao, Lihe Ding, Xin Cai, Zhanpeng Huang, Zibin Wang, and Tianfan Xue. Lora-edit: Controllable first-frame-guided video editing via mask-aware lora fine-tuning, 2025. 2

[6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[7] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2

[8] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli

Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion, 2023. 1, 2

[9] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. 2

[10] Yaron Lipman, Marton Havasi, Peter Holderrieth, Neta Shaul, Matt Le, Brian Karrer, Ricky T. Q. Chen, David Lopez-Paz, Heli Ben-Hamu, and Itai Gat. Flow matching guide and code, 2024. 2

[11] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4296–4304, 2024. 2

[12] Ziheng Ouyang, Zhen Li, and Qibin Hou. K-lora: Unlocking training-free fusion of any subject and style loras. In *CVPR*, 2025. 2

[13] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 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*, pages 22500–22510, 2023. 1, 2

[14] Viraj Shah, Nataniel Ruiz, Forrester Cole, Erika Lu, Svetlana Lazebnik, Yuanzhen Li, and Varun Jampani. Ziplora: Any subject in any style by effectively merging loras. 2023. 1, 2

[15] Vera Soboleva, Aibek Alanov, Andrey Kuznetsov, and Konstantin Sobolev. T-lora: Single image diffusion model customization without overfitting, 2025. 2, 4

[16] Gowthami Somepalli, Anubhav Gupta, Kamal Gupta, Shramay Palta, Micah Goldblum, Jonas Geiping, Abhinav Shrivastava, and Tom Goldstein. Measuring style similarity in diffusion models, 2024. 5

[17] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

[18] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction, 2024. 2

[19] Youcan Xu, Zhen Wang, Jun Xiao, Wei Liu, and Long Chen. Freetuner: Any subject in any style with training-free diffusion. *arXiv preprint arXiv:2405.14201*, 2024. 2

[20] Yang Yang, Wen Wang, Liang Peng, Chaotian Song, Yao Chen, Hengjia Li, Xiaolong Yang, Qinglin Lu, Deng Cai, Boxi Wu, and Wei Liu. Lora-composer: Leveraging low-rank adaptation for multi-concept customization in training-free diffusion models, 2024. 1, 2

[21] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *The Thirteenth International Conference on Learning Representations*, 2025. 2

[22] Yisu Zhang, Chenjie Cao, Chaohui Yu, and Jianke Zhu. Lion-lora: Rethinking lora fusion to unify controllable spatial and temporal generation for video diffusion, 2025. 2

[23] Ming Zhong, Yelong Shen, Shuohang Wang, Yadong Lu, Yizhu Jiao, Siru Ouyang, Donghan Yu, Jiawei Han, and Weizhu Chen. Multi-lora composition for image generation, 2024. 2