T-LoRA: Single Image Diffusion Model Customization Without Overfitting

Vera Soboleva[1, 2], **Aibek Alanov**[1, 2], **Andrey Kuznetsov**[1, 3], **Konstantin Sobolev**[1, 4*]

[1]FusionBrain Lab,
[2]HSE University,
[3]Innopolis University,
[4]Lomonosov Moscow State University
sobolevkv@my.msu.ru

Abstract

While diffusion model fine-tuning offers a powerful approach for customizing pre-trained models to generate specific objects, it frequently suffers from overfitting when training samples are limited, compromising both generalization capability and output diversity. This paper tackles the challenging yet most impactful task of adapting a diffusion model using just a single concept image, as single-image customization holds the greatest practical potential. We introduce *T-LoRA*, a **T**imestep-Dependent **Lo**w-**R**ank **A**daptation framework specifically designed for diffusion model personalization. In our work we show that higher diffusion timesteps are more prone to overfitting than lower ones, necessitating a timestep-sensitive fine-tuning strategy. *T-LoRA* incorporates two key innovations: (1) a dynamic fine-tuning strategy that adjusts rank-constrained updates based on diffusion timesteps, and (2) a weight parametrization technique that ensures independence between adapter components through orthogonal initialization. Extensive experiments show that *T-LoRA* and its individual components outperform standard LoRA and other diffusion model personalization techniques. They achieve a superior balance between concept fidelity and text alignment, highlighting the potential of *T-LoRA* in data-limited and resource-constrained scenarios.

Code — https://github.com/ControlGenAI/T-LoRA
Extended version — https://arxiv.org/abs/2507.05964

1 Introduction

Recent advances in personalizing large-scale text-to-image diffusion models (Ruiz et al. 2023; Gal et al. 2022; Kumari et al. 2023) have revolutionized content creation, enabling pre-trained models with the ability to generate highly specific and customized outputs, such as particular objects, styles, or domains. The primary objectives of such customization are twofold: (1) ensuring high-quality preservation of the target concept, and (2) achieving precise alignment between the generated image and the input text.

Fine-tuning based customization approaches are effective at producing high fidelity concept samples (Ruiz et al. 2023; Kumari et al. 2023). However, they encounter limitations due to the restricted dataset size, which hinder generalization and cause artifacts like background elements and pose information to "leak" into outputs, reducing the model's flexibility and creativity. Compared to standard fine-tuning, lightweight Low-Rank Adaptater (LoRA) (Hu et al. 2021) offers significant advantages (Biderman et al. 2024; Ryu et al. 2024): it significantly reduces the number of trainable parameters, making it suitable for resource-constrained settings, and it is less prone to overfitting, thereby better preserving the original model's generative capabilities.

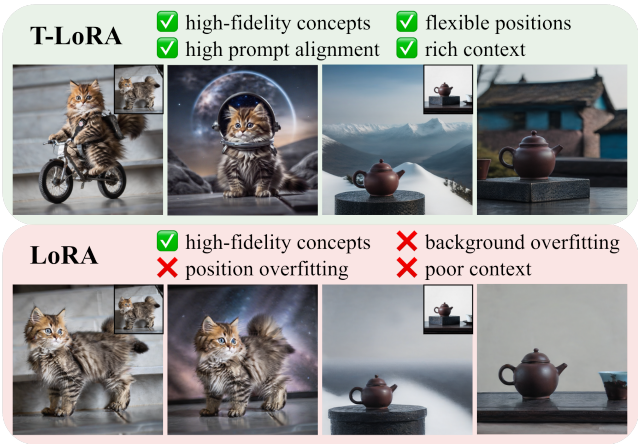


Figure 1: T-LoRA reduces overfitting related to position and background, enabling versatile and enriched generation. From left to right, the images where obtained with the following prompts: "V* riding a bike", "V* in a space helmet in a cosmic landscape", "V* on top of a snowy mountain peak", "V* with a blue house in the background".

Customizing diffusion models with a single concept image is particularly challenging due to the high risk of overfitting, even for lightweight methods (Qiu et al. 2023; Hu et al. 2021; Han et al. 2023) (Figure 1). However, this task holds significant practical value, as users often lack multiple images of their concepts with varied backgrounds, making single-image customization a key focus of our research.

We hypothesize that the root cause of overfitting lies in the fine-tuning process applied during the noisiest steps of

*Corresponding author: sobolevkv@my.msu.ru

the diffusion process. At these steps, the model is trained to recover the training images from heavily corrupted inputs, which inadvertently restricts its capacity to generate diverse and flexible scene structures. Simultaneously, these noisy steps are crucial for preserving the structural coherence and fine-grained details of the target concept. Our analysis reveals that omitting these noisy steps during fine-tuning results in a substantial loss of fidelity, underscoring the critical trade-off between maintaining concept precision and enabling generative diversity (Figure 2).

Based on our analysis, we introduce *T-LoRA*, a Timestep-Dependent Low-Rank Adaptation framework for diffusion model customization. *T-LoRA* prioritizes training capacity for less noisy timesteps while reducing signals for noisier ones using a time-dependent masking strategy, *Vanilla T-LoRA*, which restricts higher-rank LoRA components during noisier timesteps. Our further analysis reveals that standard LoRA adapters often exhibit an effective rank that is significantly smaller than the original rank hyperparameter. This property could limit the effectiveness of *Vanilla T-LoRA*, as masked and unmasked adapters may possess similar expressive power. To address this, we introduce *Ortho-LoRA*, a novel LoRA initialization technique ensuring orthogonality between adapter components, thereby explicitly separating information flows across different timesteps. Extensive experiments demonstrate the effectiveness of *T-LoRA* and its components (*Vanilla T-LoRA* and *Ortho-LoRA*). Our proposed framework significantly outperforms existing lightweight fine-tuning approaches in single-image diffusion model personalization tasks in both metrics and user study. These results highlight the potential of integrating time-dependent regularization and orthogonality into future diffusion model adaptation frameworks.

To summarize, our key contributions are as follows:

- We perform a detailed analysis of overfitting in diffusion model adaptation and reveal that it primarily occurs at higher (noisier) timesteps of the diffusion process.
- We propose *T-LoRA*, a Low-Rank Adaptation framework that mitigates overfitting in diffusion model personalization through a *rank masking strategy*, balancing training capacity and applying stronger regularization at higher timesteps.
- We explore the concept of effective rank in LoRA matrices and propose *Ortho-LoRA*, a novel *orthogonal weight initialization method*, that enhances effective rank utilization and improves information flow separation across timesteps, boosting *T-LoRA* performance.

2 Background

Diffusion models (Song, Meng, and Ermon 2020; Ho, Jain, and Abbeel 2020) are probabilistic generative models using neural networks to approximate data distributions by iteratively denoising Gaussian-sampled variables. We focus on text-to-image diffusion models ε_θ, which generate images x from text prompt P. These models use a text encoder E_T to extract text embeddings $p = E_T(P)$. Latent diffusion models (Rombach et al. 2022) encode images into latent representations $z = E(x)$ via an encoder E and decode them with a decoder D, ensuring $x \approx D(z)$. The diffusion model ε_θ is trained to predict noise in the noisy latent representation:

$$\min_\theta \mathbb{E}_{p,t,z,\varepsilon}\left[\|\varepsilon - \varepsilon_\theta(t, z_t, p)\|_2^2\right], \tag{1}$$

where $\varepsilon \sim N(0, I)$, t is the diffusion timestep, and z_t is a noisy latent code obtained via the forward process $z_t = \text{ForwardDiffusion}(t, z_0, \varepsilon)$. During inference, random noise $z_T \sim N(0, I)$ is iteratively denoised to recover z_0.

Low-Rank Adaptation (LoRA) (Hu et al. 2021) efficiently fine-tunes large pre-trained models by updating the weight matrix W as $\tilde{W} = W + BA$, where $A \in \mathbb{R}^{r\times m}$, $B \in \mathbb{R}^{n\times r}$, r is the rank, and (n, m) is the dimensionality of W. To preserve the pre-trained model's behavior initially, A is normally initialized, and B is set to zero.

Diffusion Model Customization often involves fine-tuning, where model weights are adjusted to generate specific concepts from a limited set of concept images. This aligns the model's output with user-defined concepts for tailored image generation. To associate a new concept with a special text token V^*, ε_θ is fine-tuned on a small set of concept images $\mathbb{C} = \{x\}_{i=1}^N$ using the following objective:

$$\min_\theta \mathbb{E}_{p,t,z=\mathcal{E}(x),x\in\mathbb{C},\varepsilon}\left[\|\varepsilon - \varepsilon_\theta(t, z_t, p)\|_2^2\right], \tag{2}$$

where $p = E_T(P)$ represents the text embeddings for the prompt $P =$ *"a photo of a V*"*. Fine-tuning enhancements include pseudo-token optimization (Gal et al. 2022), diffusion model fine-tuning (Ruiz et al. 2023; Kumari et al. 2023), and lightweight parameterization (Han et al. 2023; Qiu et al. 2023) to reduce costs and mitigate overfitting. LoRA (Hu et al. 2021), with its lightweight design, high concept fidelity, and strong prompt alignment, serves as a strong baseline for subject-driven generation and a key component in both fine-tuning-based (Arar et al. 2024) and encoder-based (Li et al. 2024) personalization techniques.

Overfitting Problem Existing methods often overfit to position and background, especially with limited concept images. Strategies like concept masking (Wei et al. 2023; Huang et al. 2024), prompt augmentation (Sohn et al. 2023), regularization (Ruiz et al. 2023; Kumari et al. 2023), advanced attention (Hua et al. 2023; Huang et al. 2024), and sampling (Zhou et al. 2023; Gu et al. 2024) have been explored. We identify the root cause as fine-tuning during noisiest timesteps, which reinforces background elements and limits flexibility. To address this, we propose reducing concept signals during noisy timesteps and adapting LoRA (Hu et al. 2021) to control concept injection across timesteps, enhancing generalization and diversity while mitigating overfitting.

3 Method

Motivation Previous studies have shown that different timesteps in diffusion models play distinct roles throughout the generation process (Choi et al. 2022; Deja et al. 2022; Li et al. 2023). For example, the authors of (Choi et al. 2022) categorized the behavior of diffusion timesteps into three main stages: high timesteps ($t \in [800; 1000]$) concentrate on forming coarse features, middle timesteps

Figure 2: Motivational Experiment. The figure shows fine-tuning the SD-XL model with LoRA over fixed timestep intervals. Focusing on the noisiest timesteps causes rapid overfitting, affecting positioning and backgrounds. Shifting to earlier timesteps improves text alignment and generation flexibility. However, completely excluding the noisiest timesteps ($t \in [800; 1000]$) is infeasible, as they are essential for maintaining concept fidelity.

($t \in [500; 800]$) produce perceptually rich content, and the lower timesteps ($t \in [0; 500]$) focus on removing residual noise. Additionally, works (Chang et al. 2023; Gao et al. 2023) show that high timesteps contribute to image diversity, as inadequate representation of the prompt's context during this stage makes it less likely to be restored in subsequent timesteps. Leveraging these insights, a growing body of work has introduced techniques to optimize diffusion model training. For instance, (Wang et al. 2025) proposed a time-dependent loss weighting strategy, while (Kim et al. 2025) and (Zheng et al. 2024) designed adaptive timestep sampling methods to improve the overall generation quality and diversity.

To investigate the role of different timesteps we fine-tuned the SD-XL model with LoRA over fixed timestep intervals (see Figure 2). The results show that fine-tuning at **higher timesteps** $t \in [800; 1000]$ leads to rapid overfitting, causing memorization of poses and backgrounds, which limits image diversity and prompt alignment, though these timesteps are crucial for defining shape and proportions. In the **middle timesteps** $t \in [500; 800]$, the generated context became richer, and the model better reproduce fine concept details. However, we lose information related to the overall shape. For example, the boot in Figure 2 retains fittings that correspond to the original concept, but the shoe is shorter. Finally, fine-tuning with **lower timesteps** $t \in [0; 500]$ demonstrated the best alignment with text prompts and yielded the richest generation. However, this approach struggled to accurately reproduce the intended concept, losing both shape and fine details of the object.

These findings highlight the necessity of managing the concept signal across timesteps. *Concept information injection during noisier timesteps should be limited to encourage diversity. Middle timesteps should receive more concept information to produce correct fine details. The information at the lowest timesteps does not need to be restricted, as the risk of overfitting is minimal at this stage.*

3.1 Vanilla T-LoRA

To tackle the aforementioned challenge, we propose a timestep-dependent fine-tuning strategy for diffusion models. Our method dynamically adjusts the ranks of LoRA adapters based on the diffusion timestep, allocating fewer parameters at higher timesteps and more at lower ones. This approach, called *Vanilla T-LoRA*, incorporates a masking mechanism (see Figure 3b):

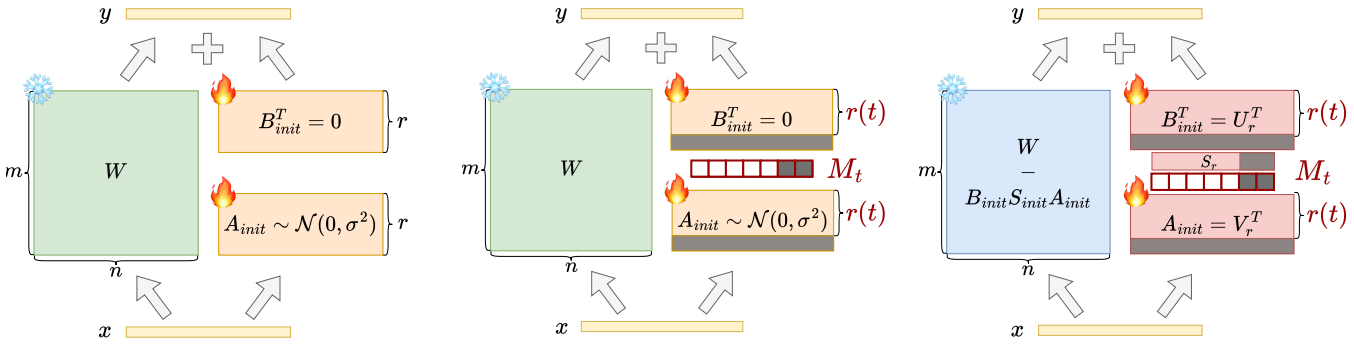
$$\tilde{W}_t = W + B_t A_t = W + B M_t A, \quad A \in \mathbb{R}^{r \times m}, B \in \mathbb{R}^{n \times r}$$
$$M_t = M_{r(t)} = \text{diag}(\underbrace{1, 1, \ldots, 1}_{r(t)}, \underbrace{0, 0, \ldots, 0}_{r - r(t)}) \in \mathbb{R}^{r \times r} \tag{3}$$

We define $r(t)$ as a linear function inversely proportional to timesteps: $r(t) = \lfloor (r - r_{\text{min}}) \cdot (T - t)/T \rfloor + r_{\text{min}}$, where r_{min} is a hyperparameter. This rank-masking strategy dynamically controls information injection across timesteps during training and inference. Higher timesteps use smaller ranks to preserve generative capabilities by focusing on coarse features and context, while lower timesteps receive more information to capture fine concept details.

3.2 On the LoRA Orthogonality

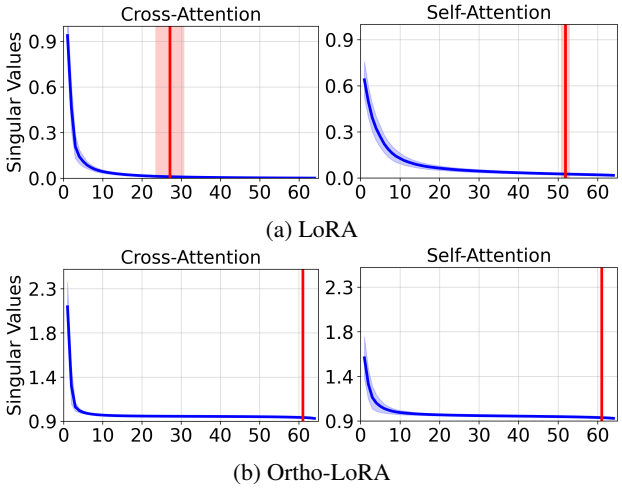
Linear dependence among the columns of LoRA matrices can compromise the effectiveness of the *Vanilla T-LoRA* masking strategy, limiting the ability to exclude information effectively. Our analysis of LoRA weights in diffusion model personalization demonstrates that their effective rank is frequently much smaller than the specified rank (Figure 4a), demonstrating linear dependency between the matrix columns. Enforcing orthogonality in A and B matrices could address this, using an SVD-like architecture and regularization, as in AdaLoRA (Zhang et al. 2023):

$$\tilde{W} = W + BSA,$$
$$L_{reg} = \lambda_{reg}(\|AA^T - I\|_F^2 + \|B^T B - I\|_F^2) \tag{4}$$



(a) LoRA (b) Vanilla T-LoRA (c) T-LoRA

Figure 3: Comparison of training methods: LoRA, the proposed Vanilla T-LoRA, and T-LoRA schemes.



(a) LoRA

(b) Ortho-LoRA

Figure 4: Singular values of B matrices for LoRA and Ortho-LoRA with $r = 64$ after 800 training steps. The red line marks the rank capturing 95% of the total singular value sum. LoRA matrices are effectively low-rank – especially in cross-attention – while Ortho-LoRA maintains full rank.

where $A \in \mathbb{R}^{r \times m}$ and $B \in \mathbb{R}^{n \times r}$ are normally initialized, and $S \in diag(\mathbb{R}^r)$ is zero-initialized. AdaLoRA demonstrated that this setup requires approximately 10,000 iterations to achieve orthogonality, which is significantly higher than the 1,000-2,000 iterations typically needed for diffusion model customization. As a result, this regularization approach is not suitable for personalization tasks, making orthogonal initialization essential. Furthermore, initializing the S matrices with zeros considerably slows down the training process (see Section 4.1 for more details).

In summary, the task requires: (1) A and B to be orthogonal from the start, and (2) S not to be zero-initialized. To address these, we use a *LoRA trick* to revise LoRA weight initialization:

$$\tilde{W} = \underbrace{W - BSA}_{\text{new weights}} + \underbrace{BSA}_{\text{LoRA}} = \hat{W} + BSA \tag{5}$$

This removes the need for zero initialization ($B = 0$), enabling arbitrary weight initialization. As shown in Figure 3c, we initialize A and B using SVD (Golub and Reinsch 1971) factor matrices to enforce orthogonality: $A_{init} = V_r^T$, $B_{init} = U_r, S_{init} = S_r$, and $S_{init} = S_r$. We term this approach *Ortho-LoRA* due to its orthogonal structure.

The choice of SVD for initialization is crucial. We examine six *Ortho-LoRA* variants: using top, middle, and bottom singular components of original weights $W \in \mathbb{R}^{n \times m}$, and of a random matrix $R \in \mathbb{R}^{n \times m}$. Top-component initialization from original weights reduces to PISSA (Meng, Wang, and Zhang 2024), but we find it suboptimal for diffusion model customization. Initializing with the last SVD components of random matrix R yields optimal results (see Section 4.1). Figure 4b illustrates that, in contrast to LoRA, *Ortho-LoRA maintains full rank throughout the entire training process without requiring any orthogonal regularization.*

3.3 T-LoRA

Finally, we introduce the complete *T-LoRA* framework (Figure 3c), combining *Vanilla T-LoRA*'s timestep-dependent rank control with *Ortho-LoRA*'s orthogonal initialization, resulting in a timestep-adaptive solution for diffusion model personalization:

$$\begin{aligned} \tilde{W} &= W - B_{init} S_{init} M_t A_{init} + BSM_t A, \\ M_t &= M_{r(t)} = \text{diag}(\underbrace{1, 1, \ldots, 1}_{r(t)}, \underbrace{0, 0, \ldots, 0}_{r - r(t)}) \end{aligned} \tag{6}$$

where $A_{init} = V^T[-r:]$, $B_{init} = U[-r:]$ and $S_{init} = S[-r:]$ are the last SVD components of a random matrix $R = USV^T, R \sim \mathcal{N}(0, 1/r)$. The rank schedule follows $r(t) = \lfloor (r - r_{min}) \cdot (T - t)/T \rfloor + r_{min}$. We do not apply any orthogonality-enforcing regularization, as the Ortho-LoRA initialization inherently maintains the orthogonality of matrices throughout the entire training process.

4 Experiments

Dataset We use 25 concepts from prior works (Gal et al. 2022; Ruiz et al. 2023; Kumari et al. 2023), including pets,

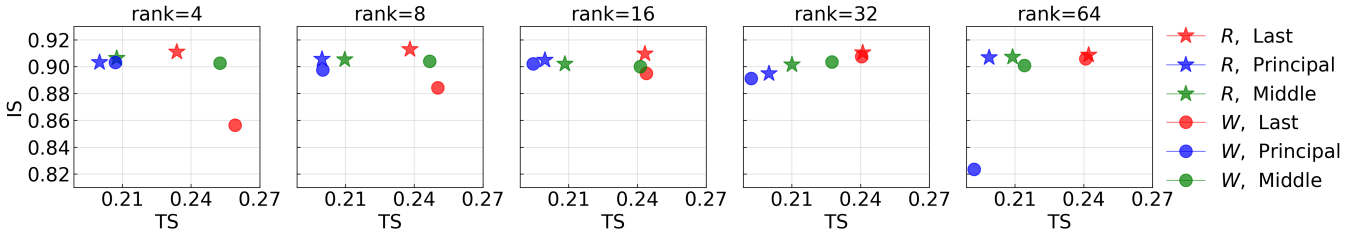


Figure 5: Results for six Ortho-LoRA initialization variants based on principal, middle, and last SVD components of the original weights W, and a random matrix R. Higher singular values correlate with overfitting, while too small values can slow down the training. Initialization with last singular values from R is optimal across most ranks.

toys, interior objects, accessories, and more. Each concept is trained on a single, manually selected image with clear visibility. For evaluation, each is paired with 25 contextual prompts (appearance, position, background changes) and 6 complex prompts (e.g., background + accessorization). We generate 5 images per contextual prompt and 15 per base prompt (e.g., "a photo of V^*"), totaling 800 concept-prompt pairs. All experiments use all 25 concepts, except the Initialization Investigation, which uses 5 randomly sampled concepts to reduce compute.

Evaluation Metric To assess concept fidelity, we compute the average pairwise cosine similarity (IS) between CLIP ViT-B/32 (Radford et al. 2021) embeddings of real and generated images, following (Gal et al. 2022). Using the neutral prompt "a photo of V^*" ensures independence from appearance, position, and accessorization changes. Backgrounds are masked to reduce bias from training image reproduction. We also report DINO-IS, computed similarly with DINO (Caron et al. 2021) embeddings. To evaluate prompt-image alignment (TS), we calculate the average cosine similarity between CLIP embeddings of the prompt and generated images.

Experimental Setup All experiments fine-tune Stable Diffusion-XL (Podell et al. 2023) with batch size 1, updating only the diffusion U-Net while keeping the text encoder fixed. Baselines follow their original setups.

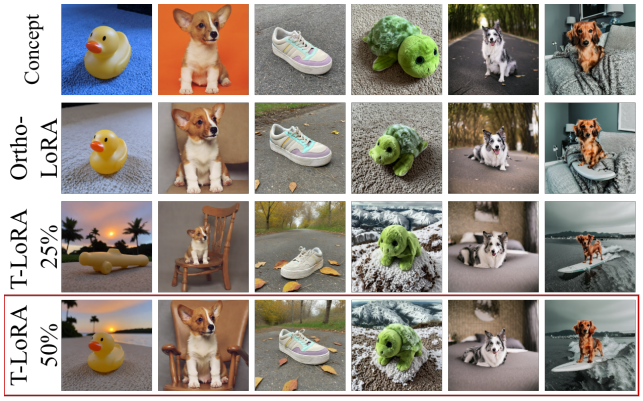


Figure 6: Generation outputs for Ortho-LoRA and T-LoRA with r_{min} set to 25% and 50% of the full rank $r = 64$. Ortho-LoRA exhibits poor alignment with the text and considerable overfitting. In contrast, T-LoRA significantly enhances alignment with the text. However, T-LoRA at 25% can struggle to accurately reproduce the concept. From left to right, the images where obtained with the following prompts: "V* with a sunset and palm trees in the background", "V* sitting on a chair", "V* with a tree and autumn leaves in the background", "V* lying on the bed", "V* riding a surfboard".

4.1 T-LoRA Design Decisions

Analysis of Ortho-LoRA Initialization Strategies We investigated six variants of the Ortho-LoRA initialization, which are based on the principal, middle, and last components of the original weights W, and a random matrix R. The IS and TS metrics for these setups are presented in Figure 5. First, we observe that for all ranks and for both R and W initializations the points in the TS metric are ordered according to the initialization singular values magnitude: the principal components initialization yields the lowest TS, followed by the middle components initialization, and the last components initialization yields the highest TS. This suggests that higher singular values are strongly correlated with overfitting. For ranks 4 and 8, initializing with the last SVD components of W turns out to be too close to zero and slows down the training process, whereas initializing with the last SVD components of R does not have this effect. Overall, initializing with the last SVD components from a random matrix R yields optimal results for most ranks, that is why we use it in all further experiments.

Selection of r_{min} In Figure 6, we present generation examples for Ortho-LoRA and T-LoRA with r_{min} set to 25% and 50% of the full rank. Both T-LoRA variants significantly improve alignment with the text and enable a greater variety of positions and backgrounds for the concepts. As r_{min} decreases, the generation becomes more flexible. However, while the 50% performs well across most concepts, the 25% is often too small, leading to reduced concept fidelity. Thus, we use T-LoRA at 50% in all subsequent experiments.

4.2 Comparison with LoRA

In Table 1, we present image and text similarity for LoRA, Vanilla T-LoRA, and T-LoRA across various ranks. For all ranks, both Vanilla T-LoRA and T-LoRA demonstrate superior text similarity compared to LoRA, while maintaining same image similarity that differs from LoRA by only a third

Methods	Rank = 4		Rank = 8		Rank = 16		Rank = 32		Rank = 64	
	IS	TS	IS	TS	IS	TS	IS	TS	IS	TS
LoRA	0.890	0.250	0.897	0.249	0.900	0.243	0.901	0.238	0.901	0.232
Vanilla T-LoRA	0.894	**0.259**	0.892	**0.261**	0.902	0.256	0.904	0.248	0.902	0.240
T-LoRA	0.899	0.255	0.897	0.260	0.897	**0.260**	0.899	**0.259**	0.900	**0.256**

Table 1: Image Similarity (IS) and Text Similarity (TS) for LoRA, Vanilla T-LoRA, and T-LoRA across dfferent ranks.

Metric	*T-LoRA-64*	LoRA-64	OFT-32	OFT-16	GSOFT-64	GSOFT-32	SVDiff
DINO-IS	0.802	**0.808**	0.804	0.802	0.806	0.804	0.414
IS	0.900	**0.901**	**0.901**	0.899	**0.901**	**0.901**	0.753
TS	0.256	0.232	0.247	0.212	0.247	0.212	**0.295**

Table 2: Image Similarity (IS) and Text Similarity (TS) for T-LoRA compared to the baseline models.

Methods	Concept Preservation		Text Alignment		Overall Preference	
	T-LoRA	Alternative	*T-LoRA*	Alternative	*T-LoRA*	Alternative
Ortho-LoRA-64	**50.3**	49.7	**58.5**	41.5	**59.3**	40.7
Vanilla T-LoRA-64	**51.7**	48.3	**60.7**	39.3	**60.3**	39.7
LoRA-64	39.3	**60.7**	**71.0**	29.0	**67.3**	32.7
OFT-32	**52.5**	47.5	**58.3**	41.7	**63.5**	36.5
GSOFT-64	49.0	**51.0**	**61.5**	38.5	**60.3**	39.7
SVDiff	**90.1**	9.9	42.1	**57.9**	**55.9**	44.1

Table 3: User study results of the pairwise comparison of T-LoRA versus other baselines.

of a decimal place. At lower ranks, Vanilla T-LoRA and T-LoRA show similar performance; however, the performance improvement of T-LoRA becomes more pronounced as the rank increases. At low ranks, LoRA approaches full rank, which is why Vanilla T-LoRA performs comparably to T-LoRA. In contrast, as the ranks increase, the effectiveness of masking in Vanilla T-LoRA diminishes, while T-LoRA continues to demonstrate its full potential.

4.3 Comparison with Baselines

In addition to LoRA (Hu et al. 2021), we compare our T-LoRA with other lightweight customization methods, including OFT (Qiu et al. 2023), GSOFT (Gorbunov et al. 2024), and SVDiff (Han et al. 2023). The results are presented in Table 2.

T-LoRA achieves the best text similarity across all methods, except for SVDiff; however, SVDiff exhibits very low image similarity and often fails to accurately represent the concept. While LoRA demonstrates the highest image similarity, it also exhibits the most significant overfitting. Notably, our method's image similarity differs from LoRA's by only a third of a decimal place.

Figure 7 showcases examples of generation for each method. T-LoRA provides greater flexibility in generation concerning position and background changes while accurately representing the concept.

4.4 Multi-image Experiments

In addition to the single-image experiments, we evaluate *T-LoRA* against LoRA (Hu et al. 2021) and OFT (Qiu et al. 2023) in the multi-image diffusion model customization. In this setting, each concept is represented by multiple images featuring diverse backgrounds. Table 4 summarizes the results for experiments conducted with 1, 2, and 3 images. For *T-LoRA* and LoRA we use $r = 64$, and $n_{blocks} = 32$ for OFT as it showed the best results in single-image setup.

Methods	# Images = 1		# Images = 2		# Images = 3	
	IS	TS	IS	TS	IS	TS
LoRA-64	0.901	0.232	0.900	0.245	0.902	0.251
OFT-32	0.901	0.247	0.901	0.261	0.901	**0.267**
T-LoRA-64	0.900	**0.256**	0.901	**0.262**	0.900	0.263

Table 4: Image Similarity (IS) and Text Similarity (TS) for multi-image Customization experiments.

T-LoRA consistently outperforms LoRA in text similarity across all image counts while achieving similar image similarity. Remarkably, *T-LoRA* trained on one image surpasses LoRA trained on two or three images. Compared to OFT, *T-LoRA* excels in the two-image scenario and performs similarly in the three-image scenario.

4.5 User Study

Finally, we conduct Human Evaluation to fully investigate our model's performance. Using an original image of the concept, a text prompt, and two generated images (one from T-LoRA and the other from an alternative method), we asked users to respond to the following questions: (1) "Which image more accurately represents the original concept?" to evaluate image similarity (2) "Which image aligns

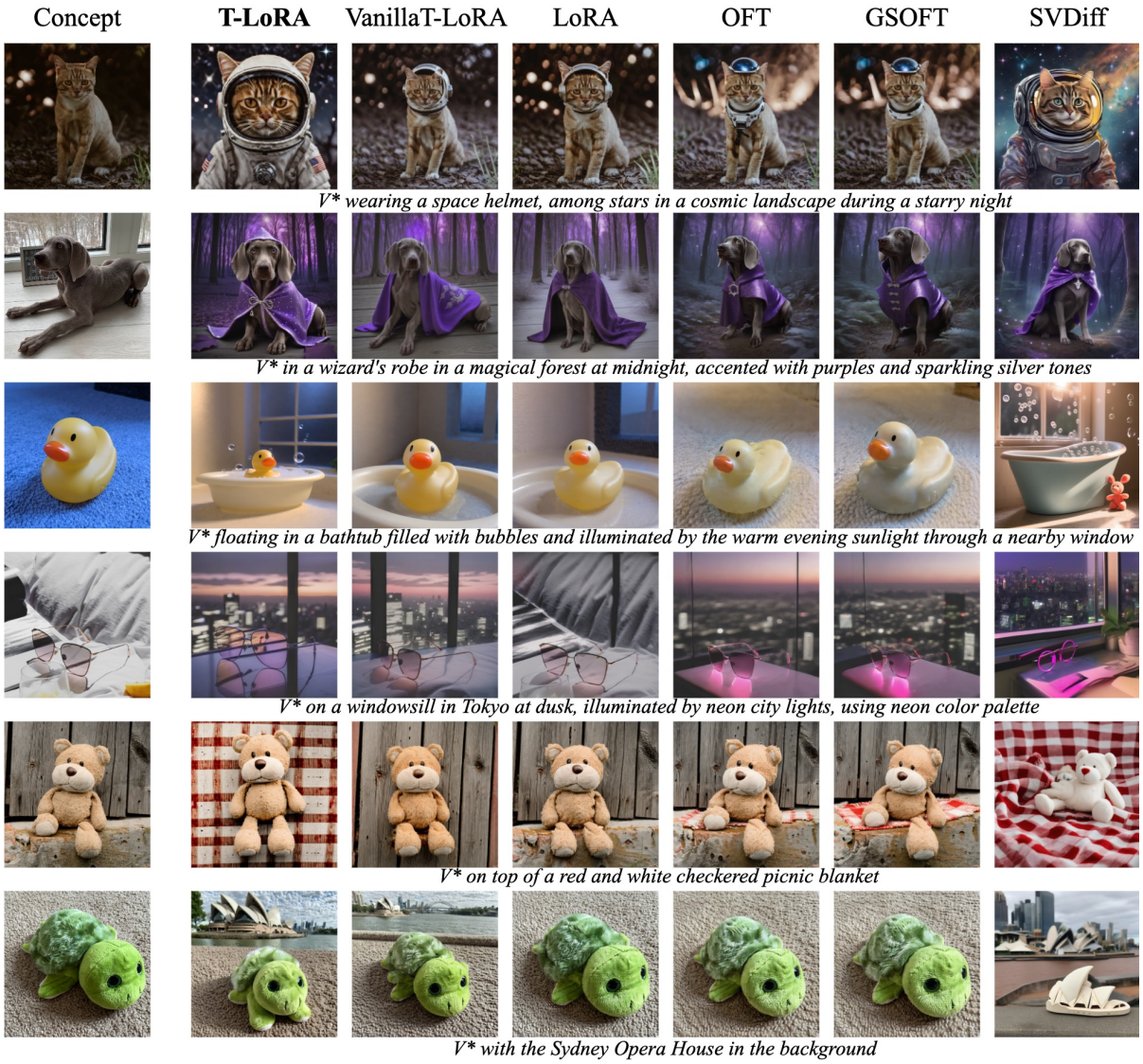


Figure 7: Generation examples for T-LoRA alongside other diffusion model customization baselines.

more closely with the text prompt?" to assess text similarity; and (3) "Which image overall demonstrates better alignment with the prompt and preserves the identity of the concept?" to evaluate the overall preference. For each pair of methods, we randomly generated 60 unique concept-prompt pairs. In total, we collect 1,800 human assessments across six pairs of methods.

Results in Table 3 show that T-LoRA significantly outperforms others in text similarity and overall preference, while achieving comparable or superior image similarity to most methods. The exception is LoRA, which surpasses T-LoRA in image similarity due to its tendency to overfit and fully reproduce the original image. Despite this, T-LoRA maintains a strong overall impression, highlighting its balanced performance across criteria.

5 Conclusion

This paper addressed the challenge of personalizing diffusion models using a single concept image, where overfitting and limited generative diversity are prevalent. We introduced *T-LoRA*, a Timestep-Dependent Low-Rank Adaptation framework featuring (1) a rank masking strategy to regulate training across diffusion timesteps and (2) *Ortho-LoRA*, an orthogonal weight initialization technique to enhance effective rank utilization. Extensive experiments show that *T-LoRA* outperforms prior methods, balancing concept fidelity and text alignment in data-limited settings. These findings lay a strong foundation for integrating timestep-sensitive strategies and orthogonality principles into future diffusion model frameworks, with promising implications for text-to-image generation and related creative tasks.

Acknowledgments

This work was supported by the The Ministry of Economic Development of the Russian Federation in accordance with the subsidy agreement (agreement identifier 000000C313925P4H0002; grant No 139-15-2025-012).

References

Arar, M.; Voynov, A.; Hertz, A.; Avrahami, O.; Fruchter, S.; Pritch, Y.; Cohen-Or, D.; and Shamir, A. 2024. PALP: prompt aligned personalization of text-to-image models. In *SIGGRAPH Asia 2024 Conference Papers*, 1–11.

Biderman, D.; Portes, J.; Ortiz, J. J. G.; Paul, M.; Greengard, P.; Jennings, C.; King, D.; Havens, S.; Chiley, V.; Frankle, J.; Blakeney, C.; and Cunningham, J. P. 2024. LoRA Learns Less and Forgets Less. *Transactions on Machine Learning Research*. Featured Certification.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9650–9660.

Chang, H.; Zhang, H.; Barber, J.; Maschinot, A.; Lezama, J.; Jiang, L.; Yang, M.-H.; Murphy, K. P.; Freeman, W. T.; Rubinstein, M.; et al. 2023. Muse: Text-To-Image Generation via Masked Generative Transformers. In *International Conference on Machine Learning*, 4055–4075. PMLR.

Choi, J.; Lee, J.; Shin, C.; Kim, S.; Kim, H.; and Yoon, S. 2022. Perception Prioritized Training of Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11472–11481.

Deja, K.; Kuzina, A.; Trzcinski, T.; and Tomczak, J. 2022. On Analyzing Generative and Denoising Capabilities of Diffusion-based Deep Generative Models. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 26218–26229. Curran Associates, Inc.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Gao, S.; Zhou, P.; Cheng, M.-M.; and Yan, S. 2023. Mdtv2: Masked diffusion transformer is a strong image synthesizer. *arXiv preprint arXiv:2303.14389*.

Golub, G. H.; and Reinsch, C. 1971. Singular value decomposition and least squares solutions. *Linear Algebra*, 134–151.

Gorbunov, M.; Yudin, K.; Soboleva, V.; Alanov, A.; Naumov, A.; and Rakhuba, M. 2024. Group and Shuffle: Efficient Structured Orthogonal Parametrization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Gu, J.; Wang, Y.; Zhao, N.; Fu, T.-J.; Xiong, W.; Liu, Q.; Zhang, Z.; Zhang, H.; Zhang, J.; Jung, H.; et al. 2024. Photoswap: Personalized subject swapping in images. *Advances in Neural Information Processing Systems*, 36.

Han, L.; Li, Y.; Zhang, H.; Milanfar, P.; Metaxas, D.; and Yang, F. 2023. Svdiff: Compact parameter space for diffusion fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7323–7334.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Hua, M.; Liu, J.; Ding, F.; Liu, W.; Wu, J.; and He, Q. 2023. Dreamtuner: Single image is enough for subject-driven generation. *arXiv preprint arXiv:2312.13691*.

Huang, M.; Mao, Z.; Liu, M.; He, Q.; and Zhang, Y. 2024. RealCustom: narrowing real text word for real-time open-domain text-to-image customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7476–7485.

Kim, M.; Ki, D.; Shim, S.-W.; and Lee, B.-J. 2025. Adaptive Non-Uniform Timestep Sampling for Accelerating Diffusion Model Training. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2513–2522.

Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; and Zhu, J.-Y. 2023. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1931–1941.

Li, L.; Li, H.; Zheng, X.; Wu, J.; Xiao, X.; Wang, R.; Zheng, M.; Pan, X.; Chao, F.; and Ji, R. 2023. Autodiffusion: Training-free optimization of time steps and architectures for automated diffusion model acceleration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7105–7114.

Li, Z.; Cao, M.; Wang, X.; Qi, Z.; Cheng, M.-M.; and Shan, Y. 2024. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8640–8650.

Meng, F.; Wang, Z.; and Zhang, M. 2024. Pissa: Principal singular values and singular vectors adaptation of large language models. *Advances in Neural Information Processing Systems*, 37: 121038–121072.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *The Twelfth International Conference on Learning Representations*.

Qiu, Z.; Liu, W.; Feng, H.; Xue, Y.; Feng, Y.; Liu, Z.; Zhang, D.; Weller, A.; and Schölkopf, B. 2023. Controlling text-to-image diffusion by orthogonal finetuning. *Advances in Neural Information Processing Systems*, 36.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. 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*, 22500–22510.

Ryu, S.; Don, H.; McCallum, L.; Friedlander, H.; Hinderliter, T.; laksjdjf; 2kpr; Khaliq, A.; Paglieri, D.; Langerman, J.; Andrews, J.; Zhang, M.; Nevarez, O.; Sikelianos, Z.; brian6091; Smith, E.; hysts; and milyiyo. 2024. Low-rank Adaptation for Fast Text-to-Image Diffusion Fine-tuning.

Sohn, K.; Ruiz, N.; Lee, K.; Chin, D. C.; Blok, I.; Chang, H.; Barber, J.; Jiang, L.; Entis, G.; Li, Y.; et al. 2023. Styledrop: Text-to-image generation in any style. *arXiv preprint arXiv:2306.00983*.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

Wang, K.; Shi, M.; Zhou, Y.; Li, Z.; Yuan, Z.; Shang, Y.; Peng, X.; Zhang, H.; and You, Y. 2025. A closer look at time steps is worthy of triple speed-up for diffusion model training. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 12934–12944.

Wei, Y.; Zhang, Y.; Ji, Z.; Bai, J.; Zhang, L.; and Zuo, W. 2023. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15943–15953.

Zhang, Q.; Chen, M.; Bukharin, A.; He, P.; Cheng, Y.; Chen, W.; and Zhao, T. 2023. ADAPTIVE BUDGET ALLOCATION FOR PARAMETER-EFFICIENT FINE-TUNING. In *11th International Conference on Learning Representations, ICLR 2023*.

Zheng, T.; Jiang, P.-T.; Wan, B.; Zhang, H.; Chen, J.; Wang, J.; and Li, B. 2024. Beta-tuned timestep diffusion model. In *European Conference on Computer Vision*, 114–130. Springer.

Zhou, Y.; Zhang, R.; Sun, T.; and Xu, J. 2023. Enhancing detail preservation for customized text-to-image generation: A regularization-free approach. *arXiv preprint arXiv:2305.13579*.