Memory-Efficient Personalization of Text-to-Image Diffusion Models via Selective Optimization Strategies

Seokeon Choi[1*] Sunghyun Park[1*] Hyoungwoo Park[1] Jeongho Kim[1,2] Sungrack Yun[1]

[1]Qualcomm AI Research[†] [2]Korea Advanced Institute of Science and Technology (KAIST)

{seokchoi, sunpar, hwoopark, jeonghok, sungrack}@qti.qualcomm.com

Abstract

Memory-efficient personalization is critical for adapting text-to-image diffusion models while preserving user privacy and operating within the limited computational resources of edge devices. To this end, we propose a selective optimization framework that adaptively chooses between backpropagation on low-resolution images (BP-low) and zeroth-order optimization on high-resolution images (ZO-high), guided by the characteristics of the diffusion process. As observed in our experiments, BP-low efficiently adapts the model to target-specific features, but suffers from structural distortions due to resolution mismatch. Conversely, ZO-high refines high-resolution details with minimal memory overhead but faces slow convergence when applied without prior adaptation. By complementing both methods, our framework leverages BP-low for effective personalization while using ZO-high to maintain structural consistency, achieving memory-efficient and high-quality fine-tuning. To maximize the efficacy of both BP-low and ZO-high, we introduce a timestep-aware probabilistic function that dynamically selects the appropriate optimization strategy based on diffusion timesteps. This function mitigates the overfitting from BP-low at high timesteps, where structural information is critical, while ensuring ZO-high is applied more effectively as training progresses. Experimental results demonstrate that our method achieves competitive performance while significantly reducing memory consumption, enabling scalable, high-quality on-device personalization without increasing inference latency.

1. Introduction

The increasing popularity of text-to-image diffusion models has fueled demand for personalized content generation, enabling applications such as avatar creation [7, 17] and object customization in novel contexts [2]. To meet this demand, training-free personalization methods [9, 15, 16, 18] have emerged, eliminating the need for test-time fine-tuning.

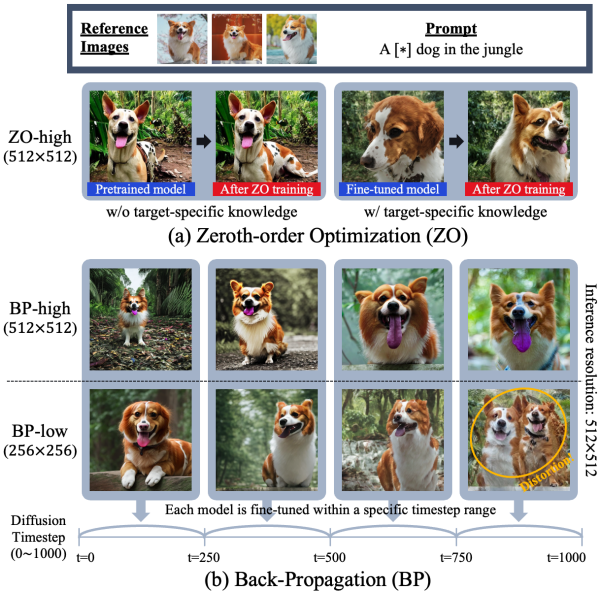

Figure 1. Key observations on different optimization strategies. (a) ZO-high requires target-specific knowledge; applying it from scratch fails to personalize the model, while using it after partial fine-tuning improves personalization. (b) Comparison of fine-tuning results across resolutions and timestep ranges using back-propagation. BP-low suffers from structural distortions due to resolution mismatch between training and inference, especially in higher timestep ranges, while BP-high achieves strong personalization but is memory-intensive and impractical for edge devices.

These approaches streamline personalization during inference; however, they often incur high memory usage, increased latency, or substantial pre-training requirements, which limit their practicality for on-device deployment.

In contrast, fine-tuning-based methods such as DreamBooth [14], Textual Inversion [4], and Custom Diffusion [8] adapt model weights using user-provided images, enabling high-fidelity personalization with fine-grained details. Due to their high computational demands, these methods are typically executed on servers. However, growing privacy concerns have increased interest in on-device personaliza-

[*] These authors contributed equally.

[†] Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.

tion, which allows users to fine-tune models locally without transmitting sensitive data. While this approach enhances privacy, the fine-tuning process remains computationally intensive and memory-demanding, posing challenges for deployment on resource-constrained edge devices.

To address this, parameter-efficient fine-tuning techniques such as LoRA [6] have been explored, which reduce the number of trainable parameters through low-rank adaptation. Although effective, LoRA still relies on full backpropagation, which requires storing activations and gradients, resulting in significant memory overhead. Zeroth-order optimization methods such as MeZO [10] avoid backpropagation by estimating gradients through forward passes, offering a more memory-efficient alternative.

However, applying MeZO directly to a pretrained model leads to poor gradient estimates and slow convergence due to the absence of target-specific cues. We observed that MeZO becomes significantly more effective when preceded by partial fine-tuning that introduces target-specific information (**Observation 1**), as shown in Fig. 1 (a). To provide such information, high-resolution backpropagation would be ideal, but it is too memory-intensive for edge devices. A practical alternative is to use low-resolution images (BP-low), which reduces memory usage by lowering activation size. Nevertheless, BP-low struggles to capture the global structure of the target object and often produces distorted outputs (**Observation 2**), particularly in the higher timestep range (750–1000), where structural information is more critical, as illustrated in Fig. 1 (b).

To overcome these limitations, we propose a selective optimization framework that combines BP-low for lightweight adaptation with zeroth-order optimization on high-resolution images (ZO-high) for structural refinement. While BP-low efficiently injects target-specific features, it tends to distort global structure at high timesteps. In contrast, ZO-high directly operates on high-resolution images, which helps preserve structural details but remains less effective early in training due to weak gradient estimates. To address this, we introduce a timestep-aware probabilistic function that dynamically selects between the two methods based on the diffusion timestep and training progress, supporting efficient personalization on edge devices without altering the standard sampling process.

Our contributions are summarized as follows:

- We propose a novel framework that combines low-resolution backpropagation and high-resolution zeroth-order optimization for memory-efficient personalization.
- We introduce a timestep-aware probabilistic function that dynamically selects the optimization strategy based on diffusion timestep and training progress.
- We demonstrate that our method enables scalable, high-quality on-device personalization with significantly reduced memory usage.

2. Proposed Method

We propose a memory-efficient optimization framework for on-device personalization of text-to-image diffusion models. Our method addresses the limitations of low-resolution backpropagation and Memory-efficient Zeroth-order Optimizer (MeZO) [10] by dynamically selecting either technique based on the diffusion timestep and training process. This hybrid strategy enables high-quality fine-tuning with low memory usage, making it suitable for edge deployment.

The proposed framework uses backpropagation on low-resolution images (BP-low) to inject target-specific information efficiently, and zeroth-order optimization on high-resolution images (ZO-high) to refine global structure with low memory overhead. A timestep-aware probabilistic function selects the method at each step, reducing BP-low overfitting and structural distortions. The full training procedure is summarized in Algorithm 1.

2.1. Overview of the Optimization Framework

Given personalization images $\mathcal{X}$ and a prompt c (*e.g.*, "a scs dog"), the goal is to fine-tune a pretrained diffusion model θ to generate images that reflect user-specific content.

At each training step i, an image x is sampled from $\mathcal{X}$, and a diffusion timestep t is drawn from a uniform distribution $t \sim \mathcal{U}(0, T)$. Based on t and i, a probabilistic function selects either BP-low or ZO-high.

2.2. Backpropagation on Low-Resolution Images

Backpropagation is applied to low-resolution images x_{low} (*e.g.*, 256×256) to reduce memory usage while capturing coarse personalization features. The optimization follows the standard gradient descent update:

$$\theta_{i+1} = \theta_i - \eta \nabla_\theta \mathcal{L}_{\text{bp}}(\theta_i; x_{\text{low}}, c, t), \tag{1}$$

where θ_i denotes the parameters at training step i, η is the learning rate, and $\mathcal{L}_{\text{bp}}$ is the diffusion reconstruction loss at timestep t.

2.3. MeZO on High-Resolution Images

We use MeZO with gradient accumulation to improve gradient approximation while maintaining memory efficiency. For each original image x (*e.g.*, 512×512), MeZO computes the accumulated gradient. For each perturbation $n \in \{1, 2, \ldots, N\}$, we sample a random perturbation vector $\boldsymbol{z}^{(n)} \sim \mathcal{N}(0, I_d)$ from a standard normal distribution and compute the gradient estimate:

$$\hat{g}_i^{(n)} = \frac{\mathcal{L}(\theta_i + \epsilon \boldsymbol{z}^{(n)}; x, c, t) - \mathcal{L}(\theta_i - \epsilon \boldsymbol{z}^{(n)}; x, c, t)}{2\epsilon} \boldsymbol{z}^{(n)}. \tag{2}$$

The final gradient estimate $\hat{g}_i$ is obtained by a weighted sum of the N estimates, $\hat{g}_i = \sum_{n=1}^{N} w_n \hat{g}_i^{(n)}$, where $w_n = \frac{1}{N}$. The parameter update is given by $\theta_{i+1} = \theta_i - \alpha \hat{g}_i$, with α as the learning rate for MeZO updates.

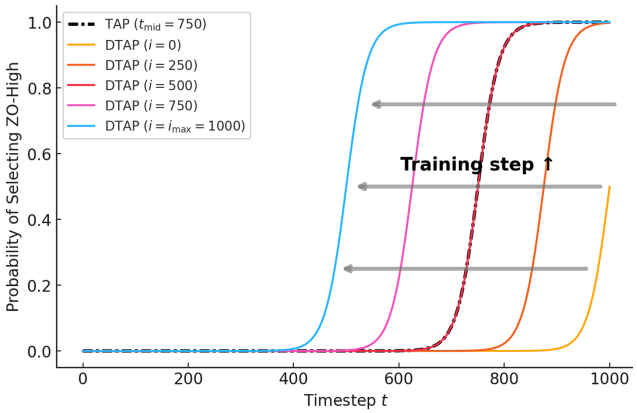

Figure 2. Comparison of TAP and DTAP over training steps when $t_{\max} = 1000$, $i_{\max} = 1000$, $k = 0.05$, and $t_{\text{mid}} = 750$. The black dashed line represents the Timestep-Aware Probability (TAP), where t_{mid} remains fixed throughout training. In contrast, the colored curves illustrate the Dynamic Timestep-Aware Probability (DTAP) at different training steps i.

2.4. Timestep-Aware Probabilistic Function

A key component of our framework is the probabilistic function that selects BP-low or ZO-high at each training step. This function adapts the optimization strategy based on the diffusion timestep t and training step i, helping to reduce overfitting and structural artifacts.

2.4.1. Timestep-Aware Probability (TAP)

Following **Observation 2**, BP-low tends to distort global structure at high timesteps ($t \approx t_{\max}$), where structural information is critical [3]. To address this, we define the probability of selecting ZO-high as:

$$p_t^{zo} = \frac{1}{1 + e^{-k(t - t_{\text{mid}})}}, \tag{3}$$

where $k > 0$ controls the steepness, and t_{mid} is the midpoint of the transition. This ensures ZO-high is more likely to be selected at higher timesteps (see Fig. 2).

2.4.2. Dynamic Timestep-Aware Probability (DTAP)

To incorporate **Observation 1**, which states that MeZO becomes more effective as training progresses, we extend TAP to dynamically adjust the midpoint over time. Specifically, we define:

$$t_{\text{dyn}}(i) = t_{\text{start}} + \frac{i}{i_{\max}}(t_{\text{end}} - t_{\text{start}}), \tag{4}$$

where $t_{\text{start}} = t_{\max}$ and $t_{\text{end}} = 2t_{\text{mid}} - t_{\max}$. At $i = 0.5 i_{\max}$, $t_{\text{dyn}}(i) = t_{\text{mid}}$, aligning DTAP with TAP. The final probability becomes:

$$p_{i,t}^{zo} = \frac{1}{1 + e^{-k(t - t_{\text{dyn}}(i))}}. \tag{5}$$

This adaptive selection enables the framework to leverage the strengths of both BP-low and ZO-high throughout training, improving personalization quality while maintaining memory efficiency.

Algorithm 1 Training Algorithm

Require: Pretrained diffusion model θ_0, personalization images $\mathcal{X}$, prompt c, total number of training steps $i_{\max}$, maximum timestep $t_{\max}$, number of perturbations N

for training step $i = 1$ to $i_{\max}$ **do**
  Sample image $x \in \mathcal{X}$
  Sample diffusion timestep $t \sim \mathcal{U}(0, t_{\max})$
  Compute $p_{i,t}^{zo}$ using Eq. 5
  if random() $> p_{i,t}^{zo}$ **then** ▷ BP-low
    Downsample x to x_{low}
    Compute gradient $\nabla_\theta \mathcal{L}_{\text{bp}}(\theta_i; x_{\text{low}}, c, t)$
    Update $\theta_{i+1} = \theta_i - \eta \nabla_\theta \mathcal{L}_{\text{bp}}$
  else ▷ ZO-high
    Use high-resolution image x
    Initialize accumulated gradient $\hat{g}_i = 0$
    for $n = 1$ to N **do**
      Sample perturbation $\boldsymbol{z}^{(n)}$
      Compute $\hat{g}_i^{(n)}$ using Eq. 2
      Accumulate $\hat{g}_i = \hat{g}_i + w_n \hat{g}_i^{(n)}$
    end for
    Update $\theta_{i+1} = \theta_i - \alpha \hat{g}_i$
  end if
end for

3. Experiments

3.1. Experimental Setup

We evaluate our method on multiple text-to-image diffusion models, including SD V1.5, SD V2.1 [13], SDXL [11], and SSD-1B [5], focusing on models suitable for edge-device deployment. All models were fine-tuned based on DreamBooth-LoRA [14], with a fixed rank of 4. We adopt DTAP as our optimization selection strategy and use the hyperparameters specified in Fig. 2. Experiments were conducted on the DreamBooth dataset [14], which includes 30 subjects and 25 prompts per subject. We report CLIP-I, CLIP-T [12], and DINO [1] scores to assess subject similarity and text-image alignment.

3.2. Quantitative Performance

We assess the scalability of our method across various diffusion models and compare it with full-resolution backpropagation. To analyze the effect of resolution, we vary the resize ratio $r \in \{0.5, 0.625, 0.75\}$. As shown in Tab. 1, our method consistently matches or outperforms BP-high while significantly reducing memory usage. Notably, on SDXL, it achieves improvements across all metrics with up to 33.69% lower memory consumption. Here, peak memory is defined as the maximum of BP and ZO memory usage during training. These results demonstrate the effectiveness of combining low-resolution backpropagation and zeroth-order optimization with a timestep-aware selection strategy.

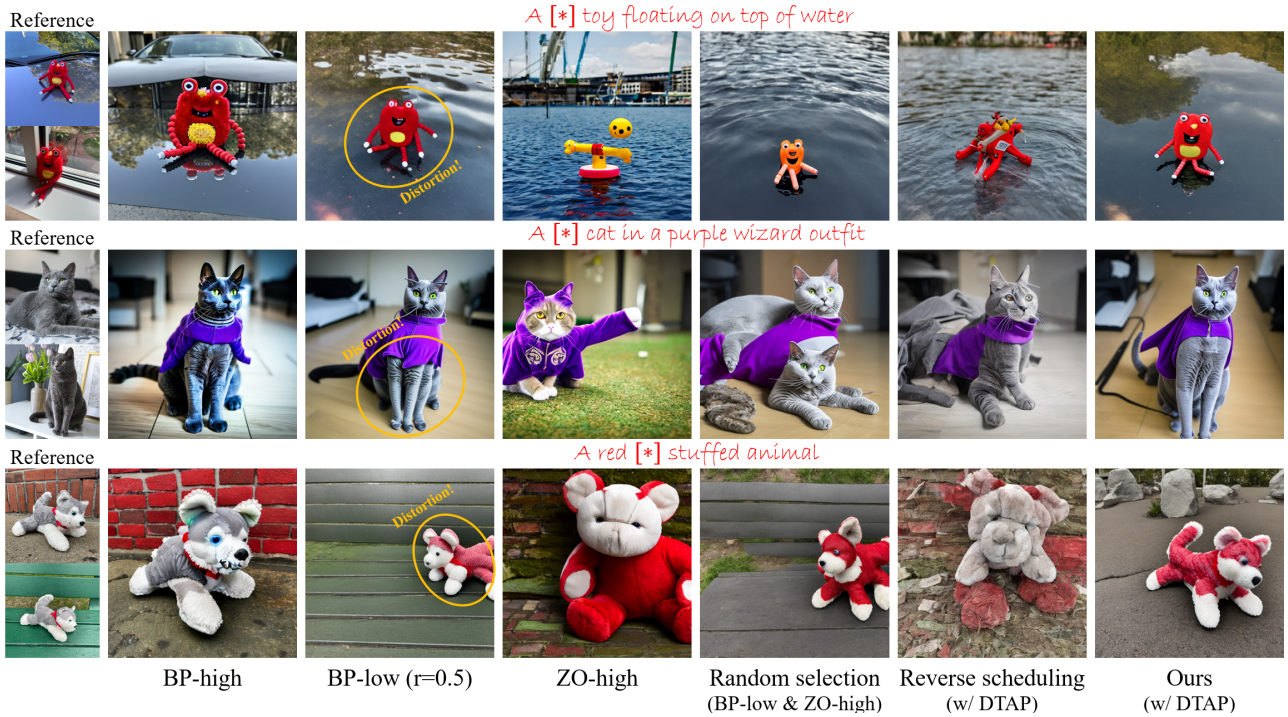

Figure 3. Qualitative comparison of optimization methods. Overfitting (BP-high), distortions (BP-low), poor personalization (ZO-high), instability (Random), and artifacts (Reversed scheduling) are observed. Our method maintains subject fidelity and structural consistency.

	Method	Performance Metrics			Mem (MiB)	
		DINO	CLIP-I	CLIP-T	BP	ZO
SD V1.5	BP-High	0.6434	**0.7926**	0.2890	5722	-
	Ours (r=0.750)	0.6403	0.7921	0.2904	5033	4756
	Ours (r=0.625)	0.6433	0.7884	**0.2914**	4737	4756
	Ours (r=0.500)	**0.6437**	0.7885	0.2904	4547	4756
SD V2.1	BP-High	0.6762	**0.8075**	0.2961	6485	-
	Ours (r=0.750)	0.6758	0.8058	**0.2969**	5832	5613
	Ours (r=0.625)	0.6787	0.8043	0.2955	5561	5613
	Ours (r=0.500)	**0.6833**	0.8067	0.2938	5368	5613
SDXL	BP-High	0.7329	0.8121	0.3002	14397	-
	Ours (r=0.750)	0.7269	0.8084	**0.3003**	11260	9546
	Ours (r=0.625)	**0.7373**	**0.8141**	0.2957	9958	9546
	Ours (r=0.500)	0.7336	0.8135	0.2941	8957	9546
SSD-1B	BP-High	0.6899	0.7955	**0.3069**	9030	-
	Ours (r=0.750)	0.6913	0.7985	0.3048	7115	7109
	Ours (r=0.625)	**0.7112**	0.8048	0.3022	6338	7109
	Ours (r=0.500)	0.7080	**0.8075**	0.2996	5729	7109

Table 1. Quantitative comparison across different diffusion models. Our method achieves comparable or superior performance to BP-high while reducing memory usage. Peak memory is defined as the maximum of BP and ZO during training and is underlined.

3.3. Qualitative Performance

Fig. 3 presents qualitative comparisons across different optimization strategies, conducted using SD 2.1. BP-low introduces structural distortions, particularly at high timesteps. ZO-high fails to capture target-specific details due to slow convergence. Random selection between BP-low and ZO-high results in unstable training and inconsistent personalization. Setting the steepness parameter k to a negative value results in a reversed scheduling, where ZO-high is applied in the early stages of training and BP-low is deferred to higher timesteps. This configuration leads to ineffective personalization and increased structural distortions, highlighting the importance of our intended timestep-aware scheduling. In contrast, our method maintains structural consistency while preserving subject-specific features. In some cases, it even surpasses BP-high in text alignment, suggesting improved generalization and reduced overfitting.

For further details, including implementation information, extended comparisons, ablation studies, hyperparameter selection, user study results, and related works, please refer to the supplementary material.

4. Conclusion

We presented a memory-efficient optimization framework for on-device personalization of text-to-image diffusion models. Rather than naively combining low-resolution backpropagation and zeroth-order optimization, our method adaptively selects between them based on key observations of diffusion timesteps. This improves personalization and structural consistency, and highlights the potential of resolution-aware training and gradient-free optimization for diffusion models. We hope this encourages further exploration of efficient personalization strategies.

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 3

[2] Daewon Chae, Nokyung Park, Jinkyu Kim, and Kimin Lee. Instructbooth: Instruction-following personalized text-to-image generation. *arXiv preprint arXiv:2312.03011*, 2023. 1

[3] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11472–11481, 2022. 3

[4] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 1

[5] Yatharth Gupta, Vishnu V Jaddipal, Harish Prabhala, Sayak Paul, and Patrick Von Platen. Progressive knowledge distillation of stable diffusion xl using layer level loss. *arXiv preprint arXiv:2401.02677*, 2024. 3

[6] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 2

[7] Liangwei Jiang, Ruida Li, Zhifeng Zhang, Shuo Fang, and Chenguang Ma. Emojidiff: Advanced facial expression control with high identity preservation in portrait generation. *arXiv preprint arXiv:2412.01254*, 2024. 1

[8] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1931–1941, 2023. 1

[9] Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36:30146–30166, 2023. 1

[10] Sadhika Malladi, Tianyu Gao, Eshaan Nichani, Alex Damian, Jason D Lee, Danqi Chen, and Sanjeev Arora. Fine-tuning language models with just forward passes. *Advances in Neural Information Processing Systems*, 36:53038–53075, 2023. 2

[11] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 3

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[13] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3

[14] 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, 3

[15] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8543–8552, 2024. 1

[16] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15943–15953, 2023. 1

[17] Yifei Zeng, Yuanxun Lu, Xinya Ji, Yao Yao, Hao Zhu, and Xun Cao. Avatarbooth: High-quality and customizable 3d human avatar generation. *arXiv preprint arXiv:2306.09864*, 2023. 1

[18] Yu Zeng, Vishal M Patel, Haochen Wang, Xun Huang, Ting-Chun Wang, Ming-Yu Liu, and Yogesh Balaji. Jedi: Joint-image diffusion models for finetuning-free personalized text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6786–6795, 2024. 1