

MobileDiffusion: Instant Text-to-Image Generation on Mobile Devices

Yang Zhao¹, Yanwu Xu, Zhisheng Xiao¹, Haolin Jia¹, and Tingbo Hou²

¹Google ²Meta GenAI
yzhao63@buffalo.edu, yanwuxu@bu.edu
zsxiao@google.com, haolinmz@google.com, houtingbo@gmail.com



Fig. 1: MobileDiffusion for (a) Text to image generation. (b) Canny edge to image, style LoRA and inpainting. Samples are all generated in one-step.

Abstract. The deployment of large-scale text-to-image diffusion models on mobile devices is impeded by their substantial model size and high latency. In this paper, we present **MobileDiffusion**, an ultra-efficient text-to-image diffusion model obtained through extensive optimizations in both architecture and sampling techniques. We conduct a comprehensive examination of model architecture design to minimize model size and FLOPs, while preserving image generation quality. Additionally, we revisit the advanced sampling technique by diffusion-GAN, and make one-step sampling compatible to downstream applications trained on the base model. Empirical studies, conducted both quantitatively and qualitatively, demonstrate the effectiveness of our proposed technologies. With them, MobileDiffusion achieves instant text-to-image generation on mobile devices, establishing a new state of the art.

1 Introduction

Text-to-image diffusion models [1, 40, 42, 44, 45, 47] have exceptional capabilities in generating high-quality images conditioned on texts. These models serve as

the foundation for a variety of applications, including image editing [15, 55], controlled generation [39, 65], personalized content generation [13, 46], video synthesis [4, 14], and low-level vision tasks [26, 61]. These large-scale models are essentially considered for necessarily running on servers with powerful neural compute units. Only a few work [25, 27] has barely touched running diffusion models on mobile devices, which remains an open challenge.

We identify two primary factors causing the inefficiency of text-to-image diffusion models. Firstly, the complexity of the network architecture in text-to-image diffusion models involves a substantial number of parameters, often reaching to the billions, resulting in computationally expensive evaluations. Secondly, the inherent design of diffusion models requires iterative denoising to generate images, necessitating multiple evaluations of the model [17, 53]. These inefficiency challenges pose a significant barrier to deploy in resource-constrained environments, such as on mobile devices. As a result, despite the potential benefits, such as enhancing user experience, addressing emerging privacy concerns, and saving cost, this aspect remains relatively unexplored within the current literature.

In this paper, we aim to develop an ultra-efficient diffusion model suitable for a range of on-device applications, necessitating a design that is lightweight, rapid, and versatile enough to handle various downstream generative tasks, including in-painting, controllable generation, and personalized generation efficiently. We approach the identified challenges—architectural and sampling efficiency—through a divide-and-conquer strategy, addressing each separately.

The challenge of **architectural efficiency** in diffusion models has been rarely addressed in existing literature. Prior attempts have focused on eliminating redundant neural network blocks [25, 27] or reorganizing them to improve efficiency [18] but lack a comprehensive analysis of the model architecture’s components. Our research fills this gap by conducting an in-depth review of diffusion networks, leading to a carefully optimized model architecture. With fewer than 400 million parameters, our model achieves high-quality image generation, surpassing previous efforts in efficiency.

The second challenge, namely the **sampling efficiency** of diffusion models, has been an active research area. With advanced numerical solvers [2, 23, 33, 34, 51] or distillation techniques [27, 35, 38, 48, 52], the necessary number of network evaluations of sampling has been reduced significantly from several hundreds to less than 10. Despite these advances, the reduced evaluation steps can still pose challenges on mobile devices, where even minimal processing demands may prove too burdensome. Our attention, therefore, turns towards pioneering single-step diffusion generation methods [49, 62, 63]. These advancements have the potential to transform diffusion models into efficient one-step text-to-image generators. Yet, there remains an unmet need for these models to flexibly accommodate a variety of conditional generation tasks through the integration of additional modules trained separately. To close this gap, we explore the nuanced design space of diffusion-GAN hybrids [28, 49, 60, 62], a forefront category of one-step diffusion models gaining popularity. Specifically, our examination centers on the UFOGen framework [62], a recent innovation in this domain. Our work advances

the discourse by conducting an in-depth analysis of UFOGen’s objective function and training methodologies, culminating in a refined approach for developing highly effective one-step diffusion GAN models. This endeavor aims to achieve a dual objective: to facilitate the creation of high-quality text-to-image synthesis and to ensure the seamless adaptability of the model for diverse downstream applications, thereby marking a significant stride towards realizing ultra-fast, versatile one-step diffusion models.

Our combined efforts in architectural design and sampling efficiency enable instant generation of high-quality 512×512 images on mobile devices, notably achieving **0.2 second** on an iPhone 15 Pro, which is about the average response time of human to visual stimulus. This achievement marks a significant leap over previous state-of-the-art in on-device text-to-image generation [27]. We introduce our model as MobileDiffusion, highlighting its potential as a foundational generative model for edge devices.

Our paper makes several key contributions to the field of on-device generative modeling, detailed as follows:

1. We present a comprehensive exploration of architectural efficiency in text-to-image diffusion models. Our work introduces a refined diffusion model architecture that is not only highly efficient and lightweight but also demonstrates rapid performance on mobile devices.
2. We study the design space of one-step diffusion-GAN models to present the best recipe for training such models, leading to high generative capabilities beyond text-to-image applications in one-single step. To our knowledge, our work pioneers the adaptation of a one-step diffusion model for a broad range of applications.
3. Synthesizing our efforts, we introduce *MobileDiffusion*, an ultra-efficient diffusion model framework specifically designed for on-device deployment.

2 Related works

As elaborated in Section 1, to improve the inference efficiency of text-to-image diffusion models and ultimately enable their deployment on mobile devices, there are two primary areas of focus: *architecture efficiency* and *sampling efficiency*. We briefly review the prior work.

Architecture Efficiency A limited number of prior works have discussed the architectural efficiency of diffusion models. Early diffusion models [10, 45, 47] adopted a UNet structure with transformer blocks. Recent works [3, 41] started to introduce vision transformers to diffusion. [18] proposed UViT, a UNet structure with a transformer backbone. With transformers establishing their advances, the differences of these methods lie in approaches to balance computation and cost, through UNet or patchify. [25] introduced an approach to distill larger diffusion models into smaller student models by selectively removing specific blocks from the teacher model. Meanwhile, [27] presented an efficient architecture search method. They trained a UNet with redundant blocks using robust training techniques [21, 64] and then pruned certain blocks based on metrics, resulting in

an architecture suitable for distillation. In our work, we align with the insights of [18] regarding the transformer components in the UNet while introducing distinctive perspectives elaborated in Section 3. In contrast to approaches like [25, 27], our focus extends beyond mere block removal. Instead, we conduct a more nuanced analysis and modification of the architecture. Additionally, unlike [25, 27], we opt not to employ knowledge distillation from a larger model. Our observations indicate that training our model from scratch yields satisfactory results.

Sampling Efficiency Diverse strategies have emerged to reduce the required sampling steps of diffusion models, broadly falling into two categories. The first involves developing fast solvers to more efficiently solve the differential equation tied to the denoising process, thereby reducing the necessary discretization steps [2, 12, 23, 33, 33]. The second approach leverages knowledge distillation techniques to compress the sampling trajectory [27, 35, 37, 48, 52]. Among them, consistency distillation [35, 36, 52] have shown promise in reducing the number of sampling steps to 4-8, while still facilitating downstream generative tasks, as highlighted by [59]. Nonetheless, attempts to further decrease the number of steps often result in diminished output quality. Recent innovations aim to revolutionize this area by enabling single-step generation [30, 49, 62, 63]. One important direction of research in one-step diffusion is diffusion-GAN hybrids [60], namely fine-tuning diffusion models with an adversarial objective [49, 62]. Despite these advances, the adaptation of one-step diffusion models for a broader array of downstream tasks remains an under-explored territory. A concurrent work [28] employs parameter-efficient adversarial fine-tuning. However, their resulting parameter-efficiently fine-tuned model does not facilitate one-step generation.

3 Designing a mobile-friendly diffusion architecture

In this section, we present our recipe for crafting highly efficient text-to-image diffusion models, which ultimately lead to sub-second generation on mobile devices. Following [42, 45], we adopt latent diffusion for its efficiency of learning text-guided generation in latent space. Our model design lies in an efficient diffusion network and a lightweight VAE decoder. We employ the CLIP-ViT/L14 [43] as the text encoder, which can run in a couple of milliseconds on-device.

3.1 Diffusion network

Early diffusion models [10, 45, 47] adopted a UNet architecture with convolutions and attentions. Recent works [3, 41] introduced vision transformer (ViT) backbones to diffusion with a great potential in scaling up the models. We noticed that transformers are expensive for a large sequence length. It needs to be considered when designing models with budget constraints. SnapFusion [27] removes transformer blocks at the highest resolution in latent space. ViT [41] patchifies a spatial input to a sequence of tokens, with a reduced length.

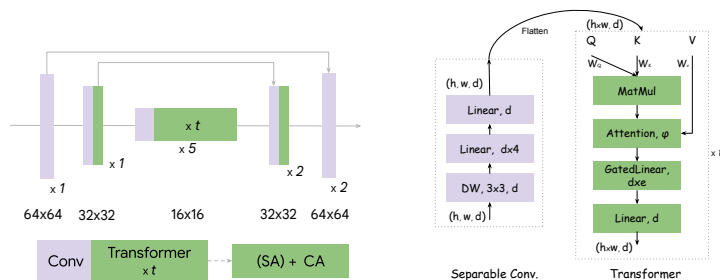


Fig. 2: Model illustration of MobileDiffusion. (Conv: Convolution. SA: Self-attention (optional). CA: Cross-attention. ϕ : Non-linearity to calculate attention weights. e : Expansion factor in gated linear layer [50]. DW: Depth-wise convolution.)

We follow the UViT [18] for designing our diffusion network with text guidance. As shown in Fig. 2, we leverage the UNet to reduce the sequence length and compute transformers efficiently. We made novel changes to the UViT architecture by conducting a comprehensive investigation of two fundamental building blocks: transformer and convolution. Throughout the study, we control the training pipeline (*e.g.* data, optimizer) to study the effects of different architectures. We optimize the network towards lower numbers of parameters and floating point operations (FLOPs), while preserving the quality. We leverage two metrics, Fréchet Inception Distance (FID) and CLIP-ViT B/32 on MS-COCO 2014 30K, complemented by visual inspections to monitor the quality of each model variation.

3.2 Optimize transformer

Scale up the backbone. Model optimization does not always indicate scaling down. By relocating model parameters, we can achieve more efficient compute. We noticed that transformers are efficient at low resolutions. Therefore, we scale up the transformer backbone in the UViT by moving more transformer layers to the bottleneck, while maintaining the total number of parameters. Additionally, the computation complexity of self-attention is $O(nd^2 + n^2d)$, where n is the sequence length, and d is the channel dimension. At low resolutions (*e.g.* 16×16), the channel dimension contributes more to the computation. Our empirical observations indicate that slightly reducing the channel dimension in the bottleneck does not adversely impact the quantitative metrics or the visual quality of the generated samples. On the other hand, attempting to stack more transformer blocks with a large reduction of channel dimension proves detrimental, resulting in an evident degradation in visual quality characterized by poor object composition and intricate artifacts. We found 1024 is the best channel dimension with a good trade-off of efficiency and quality.

Decouple SA from CA. In text-to-image diffusion models, self-attention plays a pivotal role in capturing long-range dependencies, although it comes with sig-

nificant computational costs at higher resolutions. For instance, at a resolution of 32×32 , the sequence length for self-attention is 1024. In prior work [18, 42], self-attention and cross-attention layers are moved together when approaching efficient designs. Upon further investigation, we discover that retaining cross-attention layers while discarding only the self-attention layers at high resolutions does not result in a performance drop. We conjecture that cross-attention is useful across different resolutions, as text guidance is crucial for both the global layout and local texture of the image. Notably, the computational cost of cross-attention at high resolution is significantly lower than that of self-attention, due to the smaller sequence length of text embeddings (*e.g.* 77 for SD). Consequently, removing only the self-attention layer yields a substantial efficiency boost. In light of these insights, we adopt a design that maintains performance while enhancing efficiency: entirely removing transformer blocks at the highest resolution (64×64); eliminating self-attention layers in the transformer blocks at 32×32 resolution and the outer 16×16 stack; retaining complete transformer blocks in the inner 16×16 stack and the innermost bottleneck stack.

Share key-value projections. Within an attention layer, both the key and value are projected from the same input x , denoted as $K = x \cdot W_K$ and $V = x \cdot W_V$. Our experiments reveal that adopting a parameter-sharing scheme, specifically setting $W_K = W_V$ for self-attention layers, does not adversely affect model performance. Therefore, we choose to implement this parameter-sharing strategy, resulting in approximately a 5% reduction in parameter count.

Replace gelu with swish. The GLU (Gated Linear Unit) [50] adopts the `gelu` activation function, a choice that unfortunately introduces numerical instability issues when employed in float16 or int8 inference on mobile devices due to the involvement of a cubic operation in the approximation [8]. Additionally, the `gelu` activation incurs slower computation, relying on specific hardware optimizations [54, 56]. Therefore, we propose substituting `gelu` with `swish`, which maintains a similar shape but is more cost-effective and computationally efficient. Importantly, empirical results indicate that this replacement does not lead to a degradation in metrics or perceptual quality.

Finetune softmax into relu. Given key K , query Q and value V , attention calculates $\mathbf{x} = \phi(K^\top Q)V$, where the function $\phi(\cdot)$ is the `softmax`. However, the `softmax` function, where $\text{softmax}(\mathbf{x}) = e^{\mathbf{x}} / \sum_{j=1}^N e^{x_j}$ and $\mathbf{x} = (x_0, x_1, \dots, x_N)$, is computationally expensive due to non-efficient parallelization of exponentiation and summation across the sequence length. Conversely, point-wise activations like `relu` present a quicker alternative that does not rely on specific hardware optimizations, and it can serve as a viable substitute [57]. Consequently, we propose employing `relu` in our attention computation. An intriguing finding is that there is no need to train a `relu`-attention model from scratch; instead, finetuning from a pre-trained `softmax`-attention model proves sufficient, and this fine-tuning process can be accomplished quickly, for example, within 10,000 iterations. Visual ablation results for this modification are provided in Appendix.

Models	#Channels	#ConvBlocks	#(SA+CA)	#Params(M)	#GFLOPs
SD-XL [42]	(320, 640, 1280)	17	31+31	2,600	710
SD-1.4/1.5 [45]	(320, 640, 1280, 1280)	22	16+16	862	392
SnapFusion [27]	(320, 640, 1280, 1280)	18	14+14	848	285
DiT XL/2 [41]	(1152)	0	28+0	675	525
MobileDiffusion	(320, 640, 1024)	11	15+18	386	182

Table 1: Comparison with other recognized latent diffusion models.

Trim feed-forward layers. In feed-forward layers of a transformer, the expansion ratio is default to 4, which is further doubled with gated units [50]. This amplifies the parameter count significantly. For example, a channel dimension 1280 surges to 10240 after projection. Such high dimensionality can be limiting in resource-constrained mobile applications. After thorough ablations, we found that trimming the expansion ratio to 3 results in nearly identical performance. With this adjustment, the FID score experiences only a slight increase of 0.27, while concurrently leading to 10% reduction in parameters count.

3.3 Optimize convolution

Separable convolution. Residual blocks in vanilla convolutional layers typically involve a substantial number of parameters, prompting various endeavors to enhance the parameter efficiency of convolutional layers. One proven approach in this context is separable convolution [19]. We have observed that replacing vanilla convolution layers with lightweight separable convolution layers in the deeper segments of the UNet yields similar performance. As a result, we replace all the convolutional layers in the UNet with separable convolutions, except for the outermost level. The separable convolutional block we adopted, as illustrated in Figure 2, shares similarities with ConvNeXt [31] but employs a smaller 3×3 kernel size. While we experimented with larger kernel sizes, such as 7×7 and 9×9 , we found that they did not provide additional improvements.

Prune redundant residual blocks. Convolution operations on high-resolution feature maps are especially computationally expensive, and pruning is a straightforward way to improve model efficiency. Through a comprehensive network search, we reduce the number of required residual blocks from 22 (in SD) to more efficient and streamlined 12. More specifically, we set 1 layer per block instead of 2 in SD, except for the innermost blocks. While maintaining a balance between resource consumption and model performance.

3.4 Model details

Employing the optimization techniques discussed earlier, we meticulously refine our selection of viable architecture candidates by imposing upper bounds on the numbers of parameters and FLOPs. We aim at 400 million parameters and 200 GFLOPs to achieve instant generation on devices. Notably, for the ease of

model search, we combine the innermost level of the down stack, the up stack, and the middle bottleneck into a unified module, as they share the same feature map dimension. This corresponds to the 5 layers in the middle of architecture as illustrated in Figure 2. Within this constrained search space, we systematically explore variations in the number of transformer layers and innermost channel dimensions, concentrating on the pivotal architectural parameters that significantly influence model performance. Due to the computational demands associated with training diffusion models, we prioritize optimizing these critical parameters to enhance the model’s efficiency and effectiveness. In Table 1, we present a comparative analysis of MobileDiffusion, against SD UNets, SnapFusion UNet [27] optimized for on-device, and DiT XL/2 [41]. It underscores that MobileDiffusion exhibits the highest efficiency. MD conveys more computation from convolution to attention than UNets, and embodies fewer FLOPs than pure transformer architectures.



Fig. 3: VAE reconstruction comparison among SD decoder, our decoder and our distilled decoder.

3.5 Optimize VAE

In latent diffusion, VAE converts a $H \times W \times 3$ image to a $\frac{H}{f} \times \frac{W}{f} \times c$ latent. SD [45] uses $f = 8$ and $c = 4$, while EMU [9] uses $f = 8$ and $c = 16$. Using smaller channel number c results in better compression but worse reconstruction. After an empirical study, we choose $f = 8$ and $c = 8$ for training our VAE. This modification enhances the quality of image reconstruction, as demonstrated in Figure 3. Similar to other VAEs used for latent diffusion models, our VAE is trained with a combination of loss functions, including \mathcal{L}_2 reconstruction loss, KL divergence for regularization, perceptual loss, and adversarial loss. We train the VAE with batch size 256 for 2 million iterations using the same dataset for diffusion model training. As pointed out in [27], when evaluating with a smaller number of diffusion steps, image decoder becomes a non-negligible component. To further enhance efficiency, we design a lightweight decoder architecture by pruning model *width* and *depth*. We train the lightweight decoder for 400K iterations with the encoder frozen. We remove the unnecessary KL regularization in the objective, and reduce the weight of adversarial loss. The distilled decoder leads to a significant performance boost, 3 times faster than SD’s decoder as shown in Table 4.

4 Elucidating the design space of UFOGen

In this section, we detail our studies on the ultra-fast one-step diffusion models, focusing on enhancing image quality and, crucially, enabling the integration of a broad spectrum of downstream applications into the one-step generation paradigm. Our investigation primarily concentrates on UFOGen [62], an innovative adversarial fine-tuning approach that has shown considerable promise for one-step text-to-image generation. We start with a concise overview of UFOGen, followed by a discussion on the series of ablation studies we undertook to refine its training methodology

4.1 Overview of UFOGen

UFOGen employs adversarial fine-tuning of a pre-existing diffusion model to achieve one-step generation. This process involves the one-step generator G_θ , being trained jointly with a discriminator D_ϕ with the following objective

$$\min_{\theta} \max_{\phi} \mathbb{E}_{q(x_0)q(x_{t-1}|x_0), p_\theta(x'_0)p_\theta(x'_{t-1}|x'_0)} \left[\underbrace{[\log(D_\phi(x_{t-1}, t))] + [\log(1 - D_\phi(x'_{t-1}, t))]}_{\text{adversarial loss}} + \underbrace{\lambda\gamma_t \|x_0 - x'_0\|^2}_{\text{diffusion loss}} \right], \quad (1)$$

where $q(x_0)$ is the training data distribution, $q(x_{t-1}|x_0)$ is the forward diffusion process. The generator is parameterized by generating a clean image $x'_0 = G_\theta(x_t, t) \sim p_\theta(x'_0)$, and $x'_{t-1} \sim p_\theta(x'_{t-1}|x'_0)$ is applying forward diffusion process on generator’s output x'_0 . Besides the adversarial loss, UFOGen also includes the original diffusion loss in its objective, and this is shown to be crucial for stabilizing the training. The generator and discriminator both have the same structure and initialized by the pre-trained diffusion model. The UNet discriminator aggregates logits across output locations make decisions.

UFOGen has demonstrated its efficacy in generating high-quality images from text prompts in a single step, significantly enhancing the efficiency of sampling in diffusion models. However, training strategies of UFOGen have not been well-studied. In addition, seamlessly incorporating the separately-trained modules such as LoRA [20], ControlNet [65] and T2I-adapter [39] with one-step sampling remains unexplored. While the original UFOGen study [62] showcases successful image generation from depth maps and canny edges, these outcomes are achieved through task-specific adversarial training, necessitating a separate copy for each new conditional task — far from the ideal scenario. To this end, we carefully studied the design space of UFOGen, drawing motivations from related methods. Our finalized design choice is obtained by comparing different variants on both the text-to-image generation and downstream conditional generation tasks.

4.2 Reconstruction term in the training objective

The first design aspect worth exploring is the role of the reconstruction term in UFOGen’s training objective. The formulation of UFOGen, as specified in

Equation 1, integrates both adversarial and reconstruction losses. The purpose of the adversarial loss is to enhance the visual accuracy of images produced through a singular diffusion step. Traditionally, a diffusion model’s single-step generation process is calibrated to forecast the *expected* pristine image from a given noisy input, symbolized by $\mathbb{E}[x_0|x_t]$. As a result, this often yields images that are somewhat blurry, lacking in fine details due to the inherent averaging across the data distribution. The inclusion of an adversarial loss term empowers the model to generate sharper images, particularly at elevated noise levels. Conversely, the functionality of the reconstruction loss within the training regime is more nuanced. It is postulated to act as a stabilizing force during the adversarial training phase, offering a straightforward learning directive. Beyond this, we propose that it plays a vital role in maintaining the integrity of the pre-trained diffusion model’s characteristics. Such preservation is crucial; in its absence, the one-step model risks deviating markedly from the original feature space of the pre-trained model. Such a divergence could severely limit the compatibility with downstream modules specifically tailored for the original diffusion framework.

Our conjecture suggests it might be helpful to provide enhanced regularization to encourage the preservation the pre-trained model’s characteristics, ensuring that the one-step model remains closely aligned with the foundational diffusion model’s feature landscape. To this end, we examined some alternative options to the reconstruction loss in Equation 1. The first one is a distillation loss that explicitly align the features of UFOGen’s generator to be aligned with the teacher model:

$$\mathcal{L}_{\text{distill}} = \left\| \text{sg} \left(G^{\text{teacher}}(x_t, t) \right) - x'_0 \right\|^2, \quad (2)$$

where $G_{\theta}^{\text{teacher}}$ is the pre-trained teacher model, and $\text{sg}(\cdot)$ stands for stop gradient. Note that the distill loss is similar to the regularization term in the objective of Adversarial Diffusion Distillation (ADD) [49].

More generally, we also test the EMA distillation objective, which is formalized as follows

$$\mathcal{L}_{\text{ema}} = \left\| \text{sg} \left(G_{\theta}^{\text{EMA}}(x_t, t) \right) - x'_0 \right\|^2, \quad (3)$$

where G_{θ}^{EMA} is the exponential moving average of G_{θ} . Given that G_{θ} originates from a pre-trained model, the EMA mechanism effectively retains a substantial portion of the pre-trained model’s information. This loss can also ensure that the essence of the pre-trained model is conserved, while providing more flexibility. Note that the distillation loss in Equation 2 is a special case of our loss, with the EMA decay parameter being set to 1. We conduct comprehensive ablation studies on these loss terms in Section 5.3.

4.3 Parameter-efficient adversarial fine-tuning

Current methods in adversarial diffusion fine-tuning, exemplified by UFOGen [62] and ADD [49], adjust the entire model. This extensive parameter modification can lead to significant shifts in the internal feature representation, potentially undermining the model’s ability to work effectively with downstream

tasks. Parameter-efficient fine-tuning emerges as a compelling solution, drawing from its successful application across both the visual and language fields [11,20]. Among these methods, LoRA (Low-Rank Adaptation) fine-tuning [20] stands out for its ability to refine diffusion models [36], offering a balanced approach that supports rapid sampling while maintaining task adaptability.

Motivated by the success of LCM-LoRA, we explore the potential of integrating LoRA fine-tuning within the UFOGen training framework. We apply LoRA exclusively to the generator, initializing the model weights from the pre-trained diffusion framework while setting the LoRA layers to initialize with a null effect. The discriminator is still initialized from the pre-trained diffusion model.

5 Experiments

We begin by providing an overview of the training specifics in Section 5.1, which encompass experimental setup, dataset, and the evaluation protocol. In Section 5.2, we present the primary results of our text-to-image generation including quantitative metrics, generated samples, perceptual assessments, as well as on-device benchmarks. In Section 5.4, we delve into an exploration of various mobile applications, such as the integration of controls through plugins and of personalization capabilities via LoRA finetuning.

5.1 Training Details

Dataset Our training process leverages a proprietary dataset, comprising an extensive collection of 150 million image-text pairs from public web [6]. The majority of images have a resolution greater than 256×256 , and 40 million images with a resolution over 512×512 . To prepare these images for training, we follow a consistent preprocessing methodology.

Optimization Our models are trained with the AdamW optimizer [32], configured with a learning rate of 0.001, β_1 of 0.9, β_2 of 0.999, and a weight decay of 0.01. We adopt a progressive training strategy: first, we train at 256×256 resolution with batch size 4096 for 0.75M steps to capture high-level semantics, followed by an additional 0.25M steps at a resolution of 512×512 with batch size 2048 to refine features.

Training Cost When searching and selecting superior candidates, we rely on the FID and CLIP scores reported at 30K training steps to decide the candidates to proceed. This corresponds to approximately 4 hours of computation using 32 TPUs with 16GB memory. We have designed a stopping mechanism for our experiments, which halts when the FID and CLIP scores show signs of slower improvement compared to previous candidates. This strategy works reasonable well in the process. In total, our endeavor consumes approximately 512 TPUs spanning 15 days to complete the network search.

Metrics We employed the MS-COCO dataset [29] as the primary source for our evaluations. In line with established practices, we present results for the

Models	Sampling	#Steps	FID-30K↓	CLIP↑	#Params(B)	#Data(B)
GigaGAN [22]	1-step	1	9.09	-	0.9	0.98
LAFITE [66]	1-step	1	26.94	-	0.23	0.003
DALL-E-2 [44]	DDPM	292	10.39	-	5.20	0.25
Imagen [47]	DDPM	256	7.27	-	3.60	0.45
SD [45]	DDIM	50	9.62	0.304	0.86	0.60
PIXART- α [5]	DPM	20	10.65	-	0.6	0.025
BK-SDM [24]	DDIM	50	16.54	-	0.50	-
SnapFusion [27]	Distilled	8	13.5	0.308	0.85	-
UFOGen [62]	1-step	1	12.78	0.317	0.86	0.6
MoibleDiffusion	DDIM	50	8.65	0.325	0.39	0.15
	1-step	1	11.67	0.320		

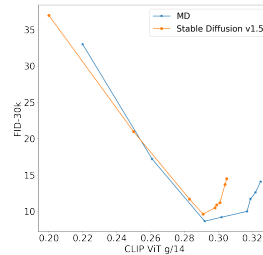
Table 2: Quantitative evaluations on zero-shot MS-COCO 2014 validation set (30K).

zero-shot FID-30K scenario, where 30,000 captions are randomly selected from the COCO validation set. These captions serve as inputs for the image synthesis process. We calculate the FID score [16] to gauge the dissimilarity between the generated samples and the 30,000 reference ground truth images. Additionally, we provide the CLIP score [43], which assesses the average similarity between the generated samples and their corresponding input captions by utilizing features extracted from a pre-trained CLIP model, OpenCLIP-ViT/g14 [7].

5.2 Text-to-Image Generation

We present comprehensive results of the efficiency and quality of text-to-image generation, spanning quantitative and qualitative comparisons, and on-device benchmarks.

Quantitative Evaluation In Table 2, we conducted a comprehensive comparison of various text-to-image generation models. For DDIM, we achieved the lowest FID scores by adjusting the *cfg* scales around 3.0. Since UFOGen does not have *cfg*, we report the FID scores as is. We also compare CLIP scores of models close to ours. While being compact in model size, our MobileDiffusion achieves better metrics than previous solutions of SnapFusion [27] and UFOGen [62]. The figure on the right shows the FID vs. CLIP curve by comparing MobileDiffusion and Stable Diffusion v1.5 using 50-step DDIM by varying *cfg*.



We also consider human preference evaluation using the standard HPS v2 benchmark [58]. We compare with SD v2.0 and v1.4 in Table 3. Despite with much smaller model size and number of inference steps, our model obtained comparable performance with Stable Diffusion variants.

On-device Benchmark We conducted benchmarking of our proposed MD using tools¹ on iPhone 15 Pro. As depicted in Table 4, MD exhibits superior

¹ <https://github.com/apple/ml-stable-diffusion/tree/main>.

Models	Anim.	Concept	Painting	Photo	Avg
SD-v2.0 (50 steps)	27.48	26.89	26.86	27.46	27.17
SD-v1.4 (50 steps)	27.26	26.61	26.66	27.27	26.95
MD (50 steps)	27.52	27.13	27.30	27.26	27.30
MD (1 step)	27.05	26.33	26.41	26.80	26.65

Table 3: HPS v2 benchmark.

efficiency in various aspects, including the text encoder, VAE decoder, UNet per-step inference, and the resulting overall latency.

Models	Text Encoder	Decoder	UNet	Steps	Overall
SD 1.5 [45]	4	285	357	20	7429
SnapFusion [27]	4	112	203	8	1740
UFOGen [62]	4	285	357	1	646
MD-UFO	4	92	142	1	238

Table 4: On-device latency (ms) measurements.

5.3 Ablation on UFOGen design choices

In this section, we ablate the design choices of UFOGen discussed in Section 4. In particular, we tried different formulations for reconstruction loss and compared LoRA fine-tuning versus full parameter fine-tuning. We report the FID scores in Table 5. Our ablation results indicate that

- LoRA adversarial fine-tuning consistently performs worse than full parameter fine-tuning
- All reconstruction losses are effective in stabilizing the adversarial fine-tuning. The distill loss obtains slightly better performance.

Reconstruction loss	No LoRA	LoRA
Diffusion loss	11.67	13.45
Distill loss	11.20	13.24
EMA distill loss	12.08	14.05

Table 5: Ablation on different adversarial finetuning settings.

Despite the slight advantage of using distill loss reflected in the FID metric, we conduct more in-depth study on the impact of training design choices on downstream application. Due to the constraint of space, we provide discussion on the study in the Appendix. We observe that, surprisingly, the original diffusion loss has the best results on applications. Therefore, our comprehensive exploration of the design space of UFOGen’s training suggest that the original setting, namely full parameter fine-tuning with diffusion loss, is the best recipe for one-step diffusion model. As a result, we adopt this setting throughout the paper.

5.4 Applications

Our MobileDiffusion framework supports a wide range of downstream conditional generation applications, as illustrated in Figure 4. On tasks including controllable generation, personalized generation and in-painting, our model consistently generate visually appealing results instantly. We want to emphasize that the conditional generation models are trained on the pre-trained diffusion models, and they are seamlessly adapted to the one-step diffusion model without any re-training. Such a flexibility enables our model to be compatible with more potential applications.

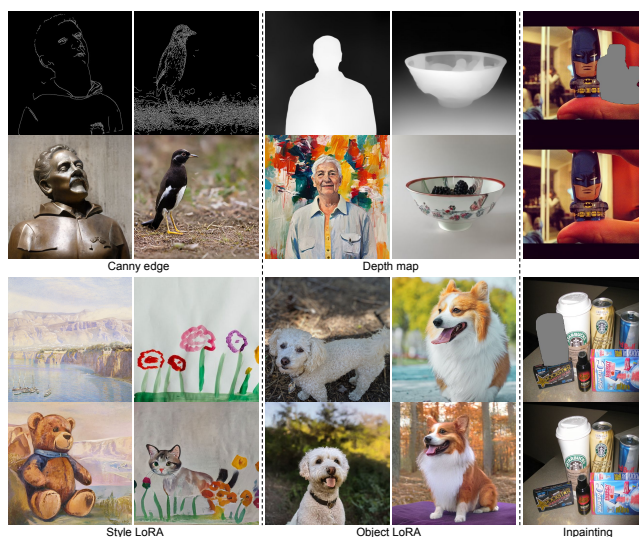


Fig. 4: MD applications on plugins, LoRA finetuning and inpainting. In each group, top row illustrates the condition or reference images and bottom row shows the results.

6 Conclusions

In this paper, we explore pushing the boundary of the diffusion model’s efficiency by proposing MobileDiffusion, with the ultimate goal of democratizing text-to-image generation on mobile devices. To achieve this goal, we conduct comprehensive studies of the architecture optimization for text-to-image diffusion models, which is a rarely touched area in prior work. Through the studies, we obtained a highly optimized architecture for diffusion UNet, with less than 400 million parameters and more efficient computational operations, while maintaining the quality. Additionally, we study the adversarial fine-tuning techniques and provide the best recipe for one-step diffusion sampling that excels in both text-to-image generation and various downstream applications. With the combined efforts, we are able to achieve the astonishing inference time of 0.2 seconds on mobile devices.

References

1. Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., et al.: ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324 (2022)
2. Bao, F., Li, C., Zhu, J., Zhang, B.: Analytic-dpm: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. arXiv preprint arXiv:2201.06503 (2022)
3. Bao, F., Nie, S., Xue, K., Cao, Y., Li, C., Su, H., Zhu, J.: All are worth words: A vit backbone for diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
4. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22563–22575 (2023)
5. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., et al.: Pixart- α : Fast training of diffusion transformer for photorealistic text-to-image synthesis. arXiv preprint arXiv:2310.00426 (2023)
6. Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyer, L., Kolesnikov, A., Puigcerver, J., Ding, N., Rong, K., Akbari, H., Mishra, G., Xue, L., Thapliyal, A., Bradbury, J., Kuo, W., Seyedhosseini, M., Jia, C., Ayan, B.K., Riquelme, C., Steiner, A., Angelova, A., Zhai, X., Houlsby, N., Soricut, R.: Pali: A jointly-scaled multilingual language-image model. In: International Conference on Learning Representations (2023)
7. Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J.: Reproducible scaling laws for contrastive language-image learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2818–2829 (2023)
8. Choi, J., Kim, M., Ahn, D., Kim, T., Kim, Y., Jo, D., Jeon, H., Kim, J.J., Kim, H.: Squeezing large-scale diffusion models for mobile. arXiv preprint arXiv:2307.01193 (2023)
9. Dai, X., Hou, J., Ma, C.Y., Tsai, S., Wang, J., Wang, R., Zhang, P., Vandenhende, S., Wang, X., Dubey, A., Yu, M., Kadian, A., Radenovic, F., Mahajan, D., Li, K., Zhao, Y., Petrovic, V., Singh, M.K., Motwani, S., Wen, Y., Song, Y., Sumbaly, R., Ramanathan, V., He, Z., Vajda, P., Parikh, D.: Emu: Enhancing image generation models using photogenic needles in a haystack. arXiv preprint arXiv:2309.15807 (2023)
10. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* (2021)
11. Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., Hu, S., Chen, Y., Chan, C.M., Chen, W., et al.: Delta tuning: A comprehensive study of parameter efficient methods for pre-trained language models. arXiv preprint arXiv:2203.06904 (2022)
12. Dockhorn, T., Vahdat, A., Kreis, K.: Genie: Higher-order denoising diffusion solvers. *Advances in Neural Information Processing Systems* **35**, 30150–30166 (2022)
13. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. In: The Eleventh International Conference on Learning Representations (2022)

14. Guo, Y., Yang, C., Rao, A., Wang, Y., Qiao, Y., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725 (2023)
15. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)
16. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* **30** (2017)
17. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Advances in neural information processing systems* **33**, 6840–6851 (2020)
18. Hoogetboom, E., Heek, J., Salimans, T.: simple diffusion: End-to-end diffusion for high resolution images. arXiv preprint arXiv:2301.11093 (2023)
19. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
20. Hu, E.J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. In: *International Conference on Learning Representations* (2021)
21. Huang, G., Sun, Y., Liu, Z., Sedra, D., Weinberger, K.Q.: Deep networks with stochastic depth. In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*. pp. 646–661. Springer (2016)
22. Kang, M., Zhu, J.Y., Zhang, R., Park, J., Shechtman, E., Paris, S., Park, T.: Scaling up gans for text-to-image synthesis. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 10124–10134 (2023)
23. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems* **35**, 26565–26577 (2022)
24. Kim, B.K., Song, H.K., Castells, T., Choi, S.: BK-SDM: Architecturally compressed stable diffusion for efficient text-to-image generation. In: *Workshop on Efficient Systems for Foundation Models @ ICML2023* (2023), <https://openreview.net/forum?id=b0VydU0XKC>
25. Kim, B.K., Song, H.K., Castells, T., Choi, S.: On architectural compression of text-to-image diffusion models. arXiv preprint arXiv:2305.15798 (2023)
26. Li, A.C., Prabhudesai, M., Duggal, S., Brown, E., Pathak, D.: Your diffusion model is secretly a zero-shot classifier. arXiv preprint arXiv:2303.16203 (2023)
27. Li, Y., Wang, H., Jin, Q., Hu, J., Chemerys, P., Fu, Y., Wang, Y., Tulyakov, S., Ren, J.: Snapfusion: Text-to-image diffusion model on mobile devices within two seconds. arXiv preprint arXiv:2306.00980 (2023)
28. Lin, S., Wang, A., Yang, X.: Sdxl-lightning: Progressive adversarial diffusion distillation. arXiv preprint arXiv:2402.13929 (2024)
29. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*. pp. 740–755. Springer (2014)
30. Liu, X., Zhang, X., Ma, J., Peng, J., Liu, Q.: InstafLOW: One step is enough for high-quality diffusion-based text-to-image generation. arXiv preprint arXiv:2309.06380 (2023)
31. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 11976–11986 (2022)

32. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2018)
33. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems* **35**, 5775–5787 (2022)
34. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095* (2022)
35. Luo, S., Tan, Y., Huang, L., Li, J., Zhao, H.: Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378* (2023)
36. Luo, S., Tan, Y., Patil, S., Gu, D., von Platen, P., Passos, A., Huang, L., Li, J., Zhao, H.: Lcm-lora: A universal stable-diffusion acceleration module. *arXiv preprint arXiv:2311.05556* (2023)
37. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. In: International Conference on Learning Representations (2021)
38. Meng, C., Rombach, R., Gao, R., Kingma, D., Ermon, S., Ho, J., Salimans, T.: On distillation of guided diffusion models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 14297–14306 (2023)
39. Mou, C., Wang, X., Xie, L., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453* (2023)
40. Nichol, A.Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Mcgrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In: *International Conference on Machine Learning*. pp. 16784–16804. PMLR (2022)
41. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 4195–4205 (2023)
42. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952* (2023)
43. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: *International conference on machine learning*. pp. 8748–8763. PMLR (2021)
44. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* **1**(2), 3 (2022)
45. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 10684–10695 (2022)
46. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: 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*. pp. 22500–22510 (2023)
47. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems* **35**, 36479–36494 (2022)

48. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. In: International Conference on Learning Representations (2022)
49. Sauer, A., Lorenz, D., Blattmann, A., Rombach, R.: Adversarial diffusion distillation. arXiv preprint arXiv:2311.17042 (2023)
50. Shazeer, N.: Glu variants improve transformer. arXiv preprint arXiv:2002.05202 (2020)
51. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2020)
52. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models (2023)
53. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: International Conference on Learning Representations (2020)
54. Sun, Z., Yu, H., Song, X., Liu, R., Yang, Y., Zhou, D.: Mobilebert: a compact task-agnostic bert for resource-limited devices. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 2158–2170 (2020)
55. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1921–1930 (2023)
56. Wang, X., Zhang, L.L., Wang, Y., Yang, M.: Towards efficient vision transformer inference: A first study of transformers on mobile devices. In: Proceedings of the 23rd Annual International Workshop on Mobile Computing Systems and Applications. pp. 1–7 (2022)
57. Wortsman, M., Lee, J., Gilmer, J., Kornblith, S.: Replacing softmax with relu in vision transformers. arXiv preprint arXiv:2309.08586 (2023)
58. Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., Li, H.: Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. arXiv preprint arXiv:2306.09341 (2023)
59. Xiao, J., Zhu, K., Zhang, H., Liu, Z., Shen, Y., Liu, Y., Fu, X., Zha, Z.J.: Ccm: Adding conditional controls to text-to-image consistency models. arXiv preprint arXiv:2312.06971 (2023)
60. Xiao, Z., Kreis, K., Vahdat, A.: Tackling the generative learning trilemma with denoising diffusion gans. In: International Conference on Learning Representations (2022)
61. Xu, J., Liu, S., Vahdat, A., Byeon, W., Wang, X., De Mello, S.: Open-vocabulary panoptic segmentation with text-to-image diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2955–2966 (2023)
62. Xu, Y., Zhao, Y., Xiao, Z., Hou, T.: Ufogen: You forward once large scale text-to-image generation via diffusion gans. arXiv preprint arXiv:2311.09257 (2023)
63. Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W.T., Park, T.: One-step diffusion with distribution matching distillation. arXiv preprint arXiv:2311.18828 (2023)
64. Yu, J., Yang, L., Xu, N., Yang, J., Huang, T.: Slimmable neural networks. In: International Conference on Learning Representations (2018)
65. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)
66. Zhou, Y., Zhang, R., Chen, C., Li, C., Tensmeyer, C., Yu, T., Gu, J., Xu, J., Sun, T.: Towards language-free training for text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17907–17917 (June 2022)