

A Architecture Optimization

Here, we provide additional complementary details that were not covered in the primary section on architecture optimization. These details are intended to offer a more comprehensive understanding of the importance and necessity of certain components.

Replace gelu with swish. As observed in Eq. 4, the function `gelu` contains a cubic term that may lead to error accumulation or overflow concerns, particularly when implementing quantization. In contrast, `swish` as shown in Eq. 5 is computationally simpler, offering a solution to mitigate potential problems.

$$\begin{aligned} \text{gelu}(x) &= x \cdot \Phi(x) \\ &\approx 0.5x(1 + \tanh(\sqrt{\frac{2}{\pi}}(x + 0.044715x^3))) \end{aligned} \quad (4)$$

where $\Phi(\cdot)$ is the cumulative distribution function for Gaussian distribution.

$$\text{swish}(x) = x \cdot \text{sigmoid}(x) \quad (5)$$

Finetune softmax into relu. Figure 5 presents a comparison of results obtained from models trained using the two activation functions. In our observations, there is rarely any noticeable visual distinction when using either of these activations in attention computations. However, employing the `relu` activation can significantly enhance mobile efficiency. As far as our knowledge extends, this represents the inaugural effort to effectively fine-tune a generative model initially activated with `softmax` into its `relu` counterpart.



A corgi's head depicted as an explosion of a nebula.

Fig. 5: Visual comparison between using `softmax` (left) and `relu` (right).

B Mobile Diffusion Architectures

In the main part, we have explained how we build efficient MD models step by step. Here we will show the detailed architectures of both MD and MD-lite in Table 6.

Models	Blocks	Layers	
		Conv	Transformer
MD	Down-1	FullConv $\times 1$	-
	Down-2	FullConv $\times 1$	CA
	Down-3	SPConv $\times 2$	$(SA \times 3 + CA \times 3) \times 2$
	Up-3	SPConv $\times 3$	$(SA \times 3 + CA \times 3) \times 3$
	Up-2	FullConv $\times 2$	CA $\times 2$
	Up-1	FullConv $\times 2$	-
MD-Lite	Down-1	FullConv $\times 1$	-
	Down-2	FullConv $\times 1$	CA
	Down-3	SPConv $\times 2$	$(SA \times 3 + CA \times 3) \times 2$
	Up-3	SPConv $\times 3$	$(SA \times 2 + CA \times 2) \times 3$
	Up-2	FullConv $\times 2$	CA $\times 2$
	Up-1	FullConv $\times 2$	-

Table 6: Architecture details of MD and MD-Lite. FullConv means original residual blocks while SPConv uses Figure 2. A notable difference between MD and MD-Lite is innermost channels which are 1024 and 896 respectively.

C Extended Generated Examples

In this section, we present extended qualitative results of text to image, adapters and LoRA finetuing.

C.1 Qualitative Examples

C.2 Text to Image

In Figure 6, we select a wide range of prompts to encompass various topics, including style, imagination, and complex compositions, beyond what was originally covered in the main section. This extended comparison underscores the consistent ability of MD to generate visually appealing results in just one-step with a highly efficient architecture.

C.3 Adapters and LoRA Finetuing

We can seamlessly adapt pretrained adapters and LoRA weights to one-step diffusion models. In Figure 8 and Figure 7, we compare qualitative results from different settings. It’s noteworthy that original UFOGen diffusion loss outperforms the other two in terms of condition and style faithfulness.

C.4 Failures Cases

In Figure 9, we illustrate two prevalent types of failures: uncommon knowledge and quantity interpretation. We attribute these failures to the limited capabilities of the text encoder, as we have observed similar phenomena in both SD-1.5 and SD-XL.

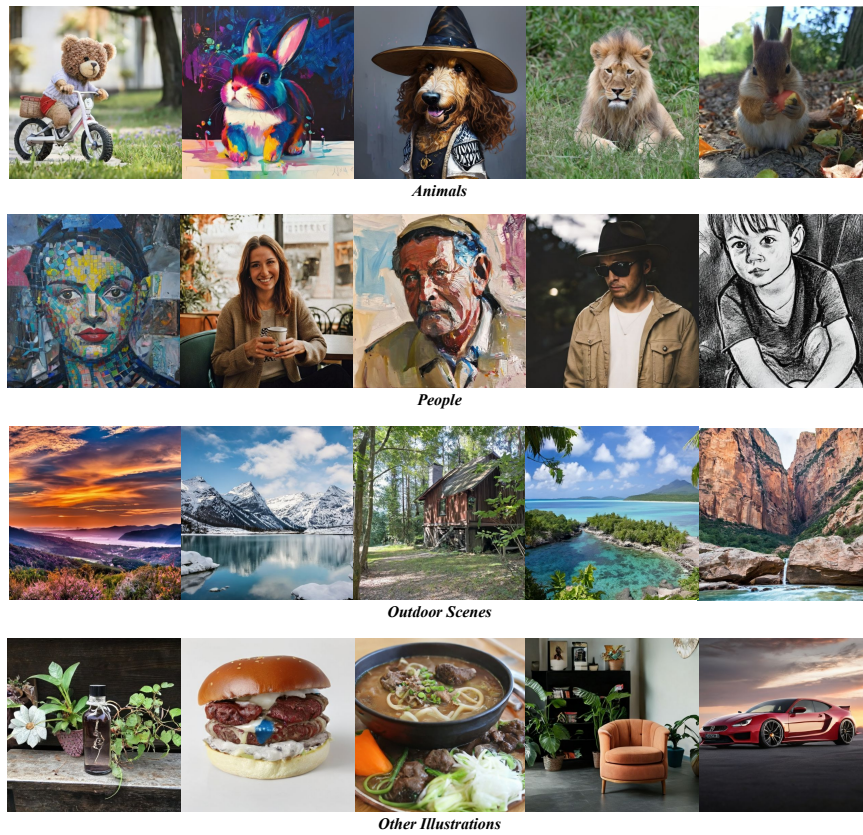


Fig. 6: Extended one-step samples from MD.







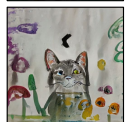
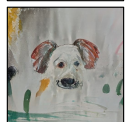


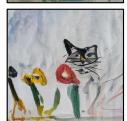
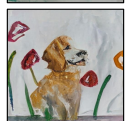






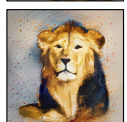

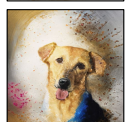
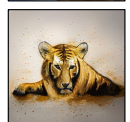
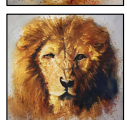
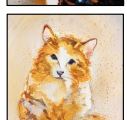
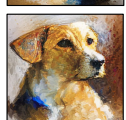
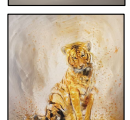
Style image	Samples				Loss
					Diffusion loss
					Distill loss
					EMA distill loss
					Diffusion loss
					Distill loss
					EMA distill loss

Fig. 7: Ablation on different adversarial finetuning settings on LoRA applications.





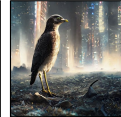



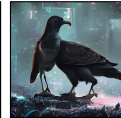



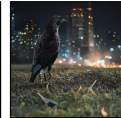
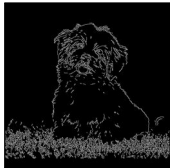



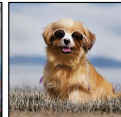








Condition image	Samples	Loss	
	   	Diffusion loss	
	   	Distill loss	
	   	EMA distill loss	
		   	Diffusion loss
		   	Distill loss
		   	EMA distill loss

Fig. 8: Ablation on different adversarial finetuning settings on adapter applications.



Fig. 9: Failure cases of MD.