EDM2+: EXPLORING EFFICIENT DIFFUSION MODEL ARCHITECTURES FOR VISUAL GENERATION

Toyota Li
Baidu Inc.
toyota.li@foxmail.com

ABSTRACT

The training and sampling of diffusion models have been exhaustively elucidated in prior art (Karras et al., 2022; 2024b). Instead, the underlying network architecture design remains on a shaky empirical footing. Furthermore, in accordance with the recent trend of scaling law, large-scale models make inroads into generative vision tasks. However, running such large diffusion models incurs a sizeable computational burden, rendering it desiderata to optimize calculations and efficiently allocate resources. To bridge these gaps, we navigate the design landscape of efficient U-Net based diffusion models, stemming from the prestigious EDM2. Our exploration route is organized along two key axes, layer placement and module interconnection. We systematically study fundamental design choices and uncover several intriguing insights for superior efficacy and efficiency. These findings culminate in our redesigned architecture, EDM2+, that reduces the computational complexity of the baseline EDM2 by $2\times$ without compromising the generation quality. Extensive experiments and comparative analyses highlight the effectiveness of our proposed network architecture, which achieves the state-of-the-art FID on the hallmark ImageNet benchmark. Code will be released upon acceptance.

1 INTRODUCTION

In recent years, diffusion models have swept the field of generative modeling, catalyzing a plethora of applications to image (Rombach et al., 2022; Podell et al., 2024; Esser et al., 2024), video (Ho et al., 2022; Blattmann et al., 2023), and 3D shape generation (Poole et al., 2023; Wang et al., 2023) in the realm of visual synthesis. Dating back to a decade ago, the advent of diffusion models relies on a plain Convolutional Neural Network (CNN) architecture (Sohl-Dickstein et al., 2015). The embarrassingly simple architecture might have posed a hindrance to the immediate blossoming of diffusion models. During the following period, one has witnessed a meteoric rise of Generative Adversarial Networks (GAN) (Goodfellow et al., 2014) in yielding photorealistic imagery (Karras et al., 2018; 2019; 2020b; 2021). In the meantime, the development of diffusion models is sluggish but has never stood still. Until 2020s, Denoising Diffusion Probabilistic Model (DDPM) (Ho et al., 2020) resurges, sparking a new wave of deep generative modeling. In this seminal work, DDPM, the introduction of U-Net (Ronneberger et al., 2015) backbone in tandem with a few modern architectural components (*e.g.*, Group Normalization (Wu & He, 2018), self-attention, and position embedding (Vaswani et al., 2017)) unleashes the potential of diffusion models in producing high-quality images comparable to other types of generative models. Henceforth, diffusion models make tremendous strides forward within the ambit of visual generative modeling.

Note that the initial adoption of U-Net in diffusion modeling borrows from other established templates, *i.e.*, its successful practice in Pixel-CNN++ (Salimans et al., 2017). Coincidentally, Jolicoeur-Martineau et al. (2021) reveal that U-Net performs substantially better than RefineNet (Lin et al., 2017) which is extensively utilized by score-based generative models (Song & Ermon, 2019; 2020). These independent observations disclose the pivotal role of network architecture design in facilitating generative modeling. In light of that, follow-up works, including iDDPM (Nichol & Dhariwal, 2021) and ADM (Dhariwal & Nichol, 2021), continuously polish the network architecture, thereby elevating the performance upper bound. These aforementioned architectures are primarily grounded on convolution operators. More recently, pure attention-based architecture further enriches the network design space, represented by Diffusion Transformer (DiT) (Peebles & Xie, 2023). The ground-

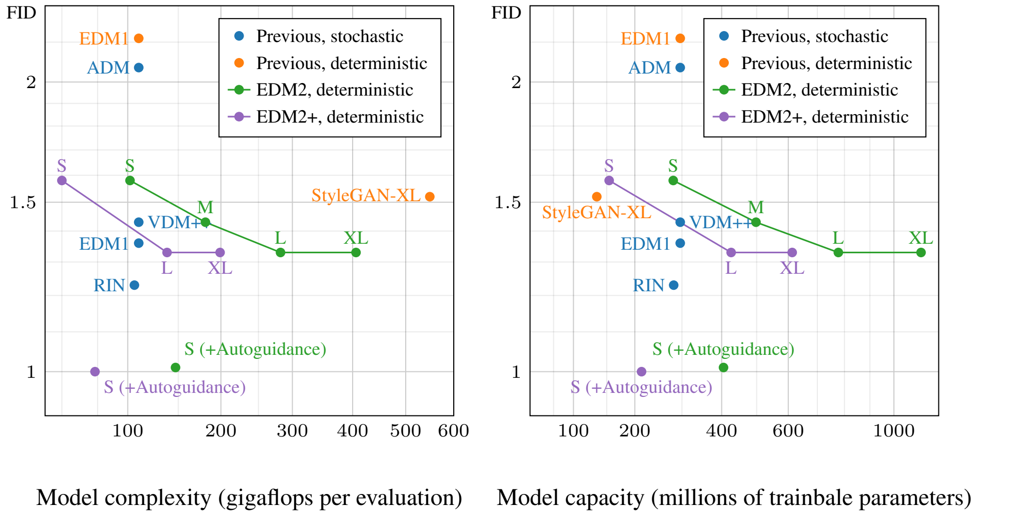

Figure 1: Our architecture EDM2+ achieves performance parity with EDM2 using 2× less compute across a wide spectrum of model sizes without guidance. Armed with the latest Autoguidance, our model is located at the bottommost leftmost corner among an array of generative architectures. In this plot, we use gigaflops per single model evaluation as a criterion of a model's intrinsic computational complexity, a similar advantage keeps consistent in terms of parameter count.

breaking Sora (Brooks et al., 2024) also adopts a spatiotemporal DiT as the foundation model for text-to-video generation. Simultaneously, Stable Diffusion 3 (Esser et al., 2024) employs a multimodal DiT (MMDiT) as the base architecture for text-to-image generation.

On the one hand, notwithstanding the flexibility and scalability of Diffusion Transformer, the final generation quality is largely dictated by its voracious appetite for the training resource. On the other hand, the top-performing diffusion models built upon relatively lightweight CNNs still prevail, *e.g.*, EDM2 (Karras et al., 2024b) showcases performance lead over Transformer-based architectures on ImageNet. Overall, EDM2 copies the U-Net macro architecture acknowledged by preceding works. Despite a few retouches of the network layers, the existing micro architecture is yet underexplored and there leaves considerable room for advanced architecture design accordingly. To fulfill this gap, we delve into the design principles and architectural choices critical for visual generation via in-depth analysis and ablation experiments. Benefiting from its streamlined and modular architecture, EDM2, as a good starting point, would ease the exploration of network design space. Startng from EDM2, we launch our investigation from the perspective of layer arrangement and inter-module connection, and eventually craft a tailored model architecture, coined as EDM2+, with on-par or even better generation quality and enhanced efficiency compared with the EDM2 counterpart.

In our design roadmap (§2), we apply the changes step-by-step to the EDM2 architecture and evaluate the impact of these individual ingredients. Our findings are in general two-fold: first, decomposing the spatial/channel mixing operations and shifting the computation focus from spatial to channel dimension strikes a better balance; second, through the lens of information bottleneck, contracting the output dimension of the embedding network concentrates the most expressive condition information to facilitate the entire information flow and naturally diminishes the parameter amount. The above conclusions are then materialized in an innovative network block, which encompasses a sequence of depthwise and pointwise convolutions, with the condition embedding sandwiched between the narrow convolution layers, as visualized in Figure 3. Our EDM2+ model architecture is comprised of dozens of such building blocks, outperforming existing top-tier diffusion models and GANs in the FID evaluation metric while remarkably reducing the model computation and storage consumption, as portrayed in Figure 1. Equipped with better utilization of guidance (Karras et al., 2024a), our endeavor sets new record FID on the ImageNet 64 × 64 benchmark, albeit using fast deterministic sampling.

Our core contributions could be summarized as:

- We conduct comprehensive experiments on the basis of EDM2 and meticulously identify the limitations in the crucial architecture components.
- We further conceptualize performance-optimized solutions, for the purpose of strengthening both the generation quality and efficiency.
- The devised architecture EDM2+ excels other leading diffusion models and GANs on the ImageNet benchmark, offering a new standard to the generative modeling field.

2 Design Roadmap

We first give a brief recap on the network design of preeminent diffusion models. Next, we provide our intuition about model design and present the roadmap to an efficient architecture while preserving the generation quality, in which the design path could be split into two branches, layer placement in the denoising network and its synergy with the embedding network.

2.1 Preliminary

EDM1 and EDM2 commonly employ the encoder-decoder paradigm in the denoising network, into which a noise-perturbed image is fed and from which a reparameterized denoised image is retrieved. Each network block stacks two 3×3 regular convolutions for feature extraction or mixing. Differently, Group Normalization (Wu & He, 2018) and SiLU nonlinearity (Ramachandran et al., 2018) precede convolution in EDM1 while EDM2 revokes such normalizations and substitutes SiLU with MP-SiLU. A skip connection is indispensable to facilitate smooth training (He et al., 2016). Formally, a network block either in the encoder or decoder of EDM2 could be formulated as

$$\boldsymbol{o} = \mathtt{conv}(\varphi(\mathtt{conv}(\varphi(\boldsymbol{x})))) + \gamma(\boldsymbol{x}), \tag{1}$$

where φ denotes the MP-SiLU activation function and γ refers to a potential linear projection that compresses channels only in the decoder part (otherwise it is identity in the encoder part). $\boldsymbol{o}$ and $\boldsymbol{x}$ enumerate the input and output feature tensor of the considered network block. Moreover, for conditional generation, a condition embedding is utilized to rectify the midway representation, written as

$$\boldsymbol{o} = \mathtt{conv}(\varphi(\mathtt{conv}(\varphi(\boldsymbol{x})) \times \phi(\boldsymbol{c}))) + \gamma(\boldsymbol{x}), \tag{2}$$

where $\boldsymbol{c}$ defines the condition information and ϕ symbolizes the mapping network that transforms it into a high-dimensional embedding. A self-attention module is suggested to append to this convolution block in case that the feature resolution is low, for example, $^1/_4$ and $^1/_8$ of the input noisy image resolution. Since self-attention does not occupy a majority of the computation in our context, we omit to discuss it in the sequel.

Reminiscent of the classical Progressive GAN (Karras et al., 2018), EDM2 draws a wealth of lessons from it: standard normal weight initialization, constant input channel concatenation, pixel normalization, and magnitude-preserving learned layers (aka equalized learning rate). EDM2 also echoes certain modern architecture design philosophies: pixel norm analogous to RMSNorm (Zhang & Sennrich, 2019), magnitude-preserving layers to weight standardization (Qiao et al., 2020) and cosine attention to QK Normalization (Dehghani et al., 2023). In turn, EDM2 now serves as a starting point for our redesigned architecture. To hit our ultimate design, we also additionally tap wisdom from other celebrated model architectures.

We shall revise the existing network details and present thorough experimental results of our exploration journey in Table 1 and 2. Concretely, the revisions are enabled one by one on top of the baseline EDM2 and we delineate the stepwise quality metric accompanied by the model parameters and computational complexity. Figure 2 and 3 sketch the architecture layout of each intermediate configuration. We perform our evaluation on the class-conditional ImageNet (Deng et al., 2009) 64×64 dataset, with the identical training recipe and data processing strategy to EDM2, in order to isolate the influence of network design. For fast prototyping, we choose a modest-sized model, EDM2-S with approximately 300M trainable parameters as the baseline (reproduced by ourselves as **CONFIG A** in Table 1), with more results for scaled-up models presented later. We follow the evaluation protocol of common practice and measure the final performance with Fréchet Inception Distance (FID) (Heusel et al., 2017) on 50,000 synthesized images (*i.e.*, FID-50K). We defer more implementation details to Section 3.

Table 1: Ablated architectures of the denoising network block. † reproduction with the official code. "c" stands for the base channel number of the first block in the entire network, while "e" the channel expansion ratio of the first pointwise convolution inside a block and "k" the kernel size of the only depthwise convolution. The same in Table 2.

Architecture	Mparams	GFLOPs	FID-50K
baseline (original publication)	280.21	101.90	1.58
A conv, c192 (baseline reproduced†)	280.21	101.90	1.63
B dwconv, c384	252.28	51.13	1.81
C dsconv, e6	273.99	72.52	1.75
D dsconv, e6, linear bottleneck	273.99	72.44	1.57
D* dsconv, e4, linear bottleneck	195.59	51.09	1.60
E mbconv, e6	273.76	72.27	1.63
E$^\diamond$ mbconv, e6, k7	278.05	75.27	1.64
F +dwconv at the end	195.76	51.26	1.68
G −dwconv in the middle	195.11	50.61	1.66

2.2 Denoising Network Design

Our overarching goal is to slim the model architecture without prejudice to the generation quality. Recall that previous lightweight network designs usually resort to a couple of depthwise and pointwise convolutions, supporting spatial and channel information mixing respectively. On this premise, we further posit that at the heart of a visual synthesis task is not only spatial pattern mixing or refinement but also composing semantically meaningful components for a high-fidelity imagery. In essence, this task is complicated by reasoning from the interaction between scene and objects, which cannot be solely reflected from superficial spatial patterns. Thus, once within a tight computational budget, we advocate trading the spatial representation mixing for stronger semantic representation learning in the channel dimension.

First, as a pilot experiment, we replace the regular convolution with depthwise convolution. Meanwhile, the channel number throughout the entire network is doubled to keep a reasonable model capacity. Then, each building block can be derived as

$$\boldsymbol{o} = \texttt{dwconv}(\varphi(\texttt{dwconv}(\varphi(\boldsymbol{x})) \times \phi(\boldsymbol{c}))) + \gamma(\boldsymbol{x}). \tag{3}$$

We observe that after roughly halving the computational complexity, the FID sacrifices not too much, which is shown as **CONFIG B** in Table 1. It indicates the relative importance of spatial and channel mixing, prompting us to allocate more computing resources to channel mixing.

Second, we speculate the optimal way to arrange the computation of channel mixing is not evenly distribute it to all layers as above. Following renowned models for efficient network design including Xception (Chollet, 2017), MobileNet series (Howard et al., 2017; Sandler et al., 2018; Howard et al., 2019; Qin et al., 2024) and EfficientNet series (Tan & Le, 2019; 2021), we substitute depthwise separable convolution (dsconv) for regular convolution, where we only expand the channel dimension in the first pointwise convolution (expansion ratio set to 6). The building block now becomes

$$\boldsymbol{o} = \texttt{pwconv}(\varphi(\texttt{dwconv}(\varphi(\texttt{pwconv}(\varphi(\texttt{dwconv}(\varphi(\boldsymbol{x})))))) \times \phi(\boldsymbol{c}))) + \gamma(\boldsymbol{x}). \tag{4}$$

This variant achieves a decent performance but is still not satisfactory, demonstrated as **CONFIG C** in Table 1. Motivated by the linear bottleneck principle in MobileNetV2 (Sandler et al., 2018), we remove the MP-SiLU in the narrow layers[1]. It also accords with the argument of "fewer activation functions" in ConvNeXt (Liu et al., 2022b). This simple modification results in the following transformation

$$\boldsymbol{o} = \texttt{pwconv}(\varphi(\texttt{dwconv}(\varphi(\texttt{pwconv}(\texttt{dwconv}(\varphi(\boldsymbol{x}))))) \times \phi(\boldsymbol{c}))) + \gamma(\boldsymbol{x}), \tag{5}$$

and effectively mitigates the representation bottleneck. Hence, **CONFIG D** in Table 1 improves the quality metric obviously compared to CONFIG C and has already surpassed the baseline CONFIG A.

[1] narrow means a small channel dimension while wide means a large one, following Sandler et al. (2018).

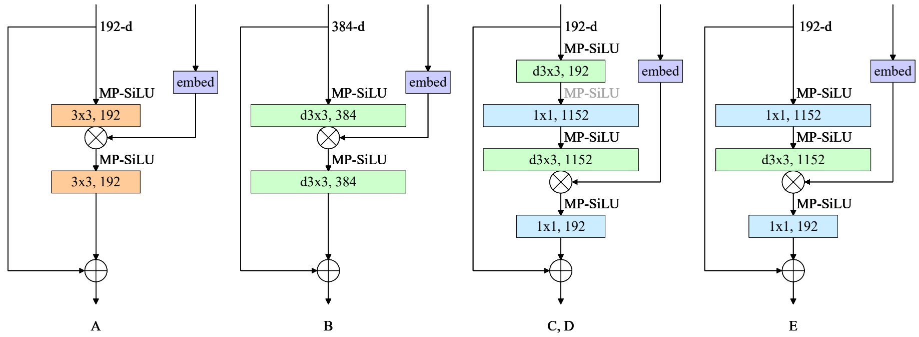

Figure 2: Block specifications of Configs A–D. A (EDM2 baseline) regular convolution. B depthwise convolution. C, D depthwise separable (depthwise + pointwise) convolution, MP-SiLU in gray indicates only existence in Config C. E mobile convolution as MobileNetV2 (Sandler et al., 2018). The width of each layer is proportional to the number of channels. Best viewed in color.

2.3 Quo Vadis, Spatial Mixing Primitives?

All the above taken into account, a question arises: to what extent do spatial mixing primitives work? To answer this question, we fade in or out depthwise convolution to scrutinize its impact.

Fade Out We erase one depthwise convolution at the start of Config C block, leaving only a single depthwise convolution and leading to a lite edition Config E,

$$\boldsymbol{o} = \texttt{pwconv}(\varphi(\texttt{dwconv}(\varphi(\texttt{pwconv}(\varphi(\boldsymbol{x})))) \times \phi(\boldsymbol{c}))) + \gamma(\boldsymbol{x}). \tag{6}$$

We find that the generation quality holds, corroborating our hypothesis that channel mixing operations outweigh spatial ones under a limited computational budget. Intriguingly, we notice that at this moment the network block looks pretty like MBConv (Tan et al., 2019), so we abuse the notation to term **CONFIG E** as mbconv in Table 1.

In addition, there appears a tendency to employ larger kernel sizes, such as 5×5 (Tan et al., 2019) or 7×7 (Liu et al., 2022b). We make an attempt to enlarge the kernel size of the only depthwise convolution to 7×7 but observe a neutral effect in terms of the quality metric, shown as **CONFIG E**$^{\diamond}$ in Table 1. This phenomenon is possibly attributed to the existence of self-attention that has already effectively captured long-range feature correspondences.

As an episode, we step a little back to scale down the channel expansion ratio of the current best-performing variant Config D from 6 to 4, so as to constrain the computational cost to nearly 50% of the baseline. This action makes the subsequent experiments more affordable and guarantees that the generation quality is still superior to the baseline, marking a promising checkpoint **CONFIG D***. We shall engage with Config D* in the remaining.

Fade In We attach another depthwise convolution at the end of Config D* block, calculated as

$$\boldsymbol{o} = \texttt{dwconv}(\varphi(\texttt{pwconv}(\varphi(\texttt{dwconv}(\varphi(\texttt{pwconv}(\texttt{dwconv}(\varphi(\boldsymbol{x}))))) \times \phi(\boldsymbol{c}))))) + \gamma(\boldsymbol{x}), \tag{7}$$

which even causes performance regression, illustrated as **CONFIG F** in Table 1.

Fade Out Given the redundancy of depthwise convolution in Config F, one of the three depthwise convolutions could be safely removed without observable drawbacks. Since eliminating the depthwise convolution at the starting position does not provide a clear gain (remember Config E *vs.*Config D), it is tentative to exclude the one in the middle, described as **CONFIG G**

$$\boldsymbol{o} = \texttt{dwconv}(\varphi(\texttt{pwconv}(\varphi(\texttt{pwconv}(\texttt{dwconv}(\varphi(\boldsymbol{x}))) \times \phi(\boldsymbol{c}))))) + \gamma(\boldsymbol{x}). \tag{8}$$

It gives rise to slightly improved performance and reduced computation overhead, as validated in Table 1.

In response to the question raised in the beginning: excessive spatial mixing primitives are indeed unnecessary for a better generation quality (Config F), while too few of them deteriorate the performance (Config E). Therefore, Config G is taken for granted in the following exploration.

Table 2: Ablated architectures of the interplay between embedding and denoising networks.

Architecture	Mparams	GFLOPs	FID-50K
A conv, c192 (baseline)	280.21	101.90	1.63
D* dsconv, e4, linear bottleneck	195.59	51.09	1.60
F +dwconv at the end	195.76	51.26	1.68
F* +dwconv at the end, embed bottleneck	154.61	51.05	1.60
G −dwconv in the middle	195.11	50.61	1.66
G* −dwconv in the middle, embed bottleneck	153.97	50.41	1.58

2.4 Embedding Network Design

The condition embedding is the crux of injecting external condition signals into the main stream of the denoising network. The condition information might be a timestamp (Ho et al., 2020) or a noise level (Song & Ermon, 2019), a class label (Dhariwal & Nichol, 2021) or a more verbose textual description (Rombach et al., 2022). In this work, we operate on the noise level and the class label information, since our exploration is set out on the ImageNet benchmark for class-conditioned image synthesis. Typically, an input numeral, representative of the condition information, is appropriately scaled and mapped to a high-dimensional space using non-learnable Fourier feature (Tancik et al., 2020) or sinusoidal embedding, in concert with a (few) learnable linear projection layer(s). This stack of neural layers is collectively dubbed as the embedding network.

It is trendy that the embedding network is gradually minimized, exemplified by the shallower mapping network in StyleGAN2-ADA (Karras et al., 2020a) or StyleGAN3 (Karras et al., 2021) and the trimmed embedding network in EDM2 (Karras et al., 2024b). The model capacity of this tiny network is presumably sufficient to extract semantic information from a single scalar condition, while a shorter path here permits the denoising network to be better informed of the condition information. Provided the network depth is extremely truncated by design, we are particularly interested in how to maximize its cooperation with the denoising network from other factors, for instance, whether the network width of their junction makes a difference to the generation quality and parameter efficiency.

Regarding the above Configs F–G, the condition embedding is mixed with a *wide* feature map. The conventional wisdom is that the condition embedding is responsible for steering the style of synthesized images in a global manner (Huang & Belongie, 2017; Karras et al., 2019). From the viewpoint of information bottleneck theory, integrating such condition information into a *narrow* feature map would be more effective, yielding more targeted and predictable control of the entire information flow. To substantiate our hypothesis, we reposition the embedding network after the last pointwise convolution, constructing an "embed bottleneck" in **CONFIG G*** (similar for **CONFIG F***)

$$\boldsymbol{o} = \texttt{dwconv}(\varphi(\texttt{pwconv}(\varphi(\texttt{pwconv}(\texttt{dwconv}(\varphi(\boldsymbol{x}))))) \times \phi(\boldsymbol{c}))) + \gamma(\boldsymbol{x}). \tag{9}$$

Notably, this step shoots two hawks with one arrow, not only reducing the network parameters by over 20% but also meliorating the generation quality. Thanks to the boosted bottleneck representation, the FID metric is again elevated beyond the baseline Config A, as exhibited in Table 2 Config F* and Config G*.

In a nutshell, the key to both efficacy and efficiency is modulating the condition embedding into the denoising network's bottleneck layers. A proof by contradiction occurs in Config E, where there are two optional positions for embedding modulation, that are, following the first pointwise convolution or the first depthwise convolution. Indeed, we do not find a noticeable difference between them (with the same FID of 1.63). It is supposed that both feature maps are of equal width, without a bottleneck of information flow in the mainstream denoising network.

3 Experiments

3.1 Datasets

We adopt ImageNet (Deng et al., 2009) pixel-space diffusion at 64×64 resolution as the benchmark. Before training, we follow ADM (Dhariwal & Nichol, 2021) protocol to pre-process the raw ImageNet dataset for a fair comparison to previous works. Specifically, the images are resized along the

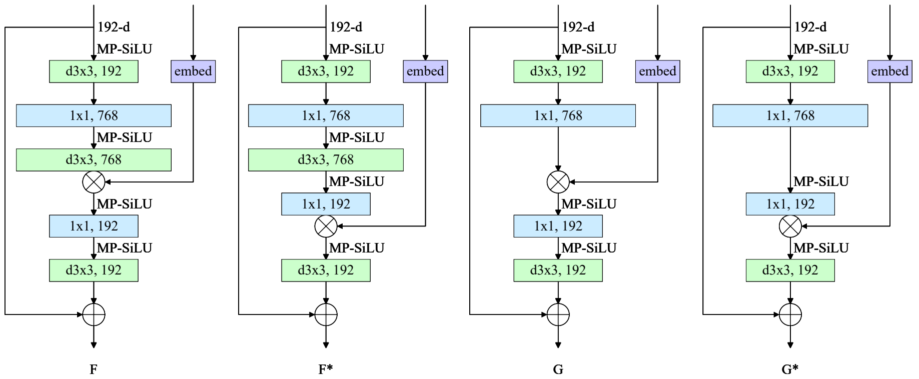

Figure 3: Block specifications of CONFIGS F–G. CONFIGS F*–G* rewire the embedding network to a narrow feature tensor in the denoising network, shaping an "embed bottleneck". G* ours EDM2+.

short edge and then cropped at the center to a desired square shape. No data augmentation is applied during training, since a large-scale dataset like ImageNet is deemed to be challenging enough to fit for most visual generation models.

3.2 EVALUATION

The evaluated checkpoints are constructed post-hoc through power-function Exponential Moving Average (EMA) over a group of snapshots with the recommended length by EDM2. The evaluation metric is the widely recognized FID (Heusel et al., 2017). It compares the distribution statistics of 50K synthesized samples against all the 1,281,167 real images in the training dataset, in line with common practice. The class labels for the 50K synthesized images are drawn from a uniform distribution. For feature extraction, we use the pre-trained Inception-v3 (Szegedy et al., 2016) model provided by StyleGAN3 (Karras et al., 2021). Limited by the computational resource, we compute FID only once, which may even put us at a disadvantage in comparison to EDM2 (because EDM2 computes FID three times and reports the *minimum*).

3.3 IMPLEMENTATION DETAILS

We implemented our network architecture based on the PyTorch (Paszke et al., 2019) library and EDM2 codebase[2]. All training runs are conducted on 32 NVIDIA A100-SMX4-80G GPUs, while each evaluation run is executed on a single node with 8 GPUs. The entire training traverses either 2147.5M or 671.1M images with a mini-batch size of 64 per device. We adopt the Adam (Kingma & Ba, 2015) optimizer with a peak learning rate of ∼0.01 and constant betas $\beta_1 = 0.9, \beta_2 = 0.99$. The learning rate is linearly warmed up over the first 10M images and decayed after 70K training iterations following a reciprocal square root schedule (Zhai et al., 2022). Larger models enjoy a moderately lower learning rate and higher dropout rate. Mixed-precision training (Micikevicius et al., 2018) is allowed to take full advantage of the tensor cores in NVIDIA Ampere architecture. Almost all activation values are cast to the 16-bit floating point (FP16) format during network forward/backward. To avoid the risk of under/overflows, it is sufficient to only cast the `NaN` and `Inf` gradient values to zeros. The second-order Heun sampler is adopted for ODE sampling, with all the hyperparameters aligned with the original EDM1 setup. The EDM2+-S model is built upon network blocks of **CONFIG G*** in Figure 3. The L-sized and XL-sized versions are obtained by scaling up the network width of EDM2+-S to 320 and 384 respectively.

3.4 QUANTITATIVE RESULTS

Comparison to deterministic sampling. As depicted in Table 3, under the scenario of deterministic sampling without guidance, we rival the generation quality of prior art diffusion models, EDM2. Of note is that the on-par quality is acquired with merely half of the computational load and parameter count. In a horizontal comparison to the GAN family with deterministic sampling, Inception-v3 based FID measurement is blamed for unfairly favoring GANs rather than diffusion models (Stein

[2] https://github.com/NVlabs/edm2

Table 3: State-of-the-art comparison on ImageNet at 64×64 resolution. NFE states the Number that the score Function is Evaluated to synthesize a single image. ↓ hints lower is better. GFLOPs tell the floating-point operations per function call.

Architecture	Deterministic		Stochastic		Model size	
	FID↓	NFE	FID↓	NFE	Mparams	Gflops
ADM (Dhariwal & Nichol, 2021)	–	–	2.07	250	296	110
+ EDM1 sampling (Karras et al., 2022)	2.66	79	1.57	511	296	110
+ EDM1 training (Karras et al., 2022)	2.22	79	1.36	511	296	110
VDM++ (Kingma & Gao, 2023)	–	–	1.43	511	296	110
RIN (Jabri et al., 2023)	–	–	1.23	1000	281	106
StyleGAN-XL (Sauer et al., 2022)	1.52	1	–	–	134	549
EDM2-S (Karras et al., 2024b)	1.58	63	–	–	280	102
+ Autoguidance(XS, $^{T}/_{8}$) (Karras et al., 2024a)	1.01	63	–	–	405	147
EDM2-M (Karras et al., 2024b)	1.43	63	–	–	498	181
EDM2-L (Karras et al., 2024b)	1.33	63	–	–	777	282
EDM2-XL (Karras et al., 2024b)	1.33	63	–	–	1119	406
EDM2+-S	1.58	63	–	–	154	50
+ Autoguidance(XS, $^{T}/_{8}$) (Karras et al., 2024a)	1.00	63	–	–	213	73
EDM2+-L	1.33	63	–	–	426	138
EDM2+-XL	1.33	63	–	–	613	199

Table 4: Runtime and memory profiling. CPU latency is timed with a batch size of 1 while the GPU throughput with a batch size of 32. The GPU throughput is measured in Frames Per Second (FPS).

Architecture	CPU Latency (s)↓	GPU Throughput (img/s)↑	GPU Memory (MB)↓
EDM2	2.069	642	1273
EDM2+	0.998 (−52%)	836 (+30%)	1114 (−13%)

et al., 2023). Therefore, previous diffusion models have to exchange more than double parameters for lower FID-50K than the best-in-class StyleGAN-XL. Still with better FID, EDM2+, among the diffusion models, is the first to preserve the same magnitude of model size as StyleGAN-XL.

Comparison to stochastic sampling. Although stochastic sampling is still at the forefront of cutting-edge diffusion models, it suffers from laborious parameter tuning and a cumbersome sampling trajectory. To outperform EDM1 using stochastic sampling, EDM2-L using deterministic sampling spends nearly triple FLOPs on a single model evaluation, partially canceling out the benefit of fewer sampling steps, while our EDM2+-L could limit the single model FLOPs to the same level as EDM1. In lieu of stochastic sampling, some concurrent works push the performance frontier of deterministic sampling with advanced Classifier-Free Guidance (CFG) (Ho & Salimans, 2021) techniques, such as guidance interval (Kynkäänniemi et al., 2024) and autoguidance (Karras et al., 2024a). Now that this work on neural architecture design is orthogonal to them, our generation performance is ready to be further improved using these complementary tricks. As a consequence, the combination of EDM2+ and autoguidance secures an FID of 1.00, reaching the state-of-the-art performance. With evidently fewer sampling steps and total compute, we are able to beat the record FID of 1.23 achieved by stochastic sampling of RIN.

Runtime analysis. Figure 1 and Table 3 mainly quantifies the model cost using FLOPs. Nevertheless, it is more practical to inspect the wall clock runtime (Ma et al., 2018). The model inference is profiled on an NVIDIA A100 GPU after warmup, with `torch.compile` and TensorFloat32 (TF32) tensor cores enabled, as well as on an Intel Xeon Platinum 8468V CPU. An apple-to-apple comparison to EDM2 on runtime and memory is collected in Table 4. The hardware execution speed is of great interest, that is improved by 52% on CPU and 30% on GPU device with our EDM2+. As a byproduct, the GPU memory volume during inference time is also shrunk by 13%.

4 Conclusion

This work invests effort into the rapidly evolving arena of diffusion model architectures, via undertaking a systematic exploration and unraveling practical guidelines for efficient network design. The valuable discoveries, pinpointing the significance of layer placement and module interconnection, are leveraged to deliver a model family named EDM2+. Our presented EDM2+ architecture achieves pronounced efficiency gains against the EDM2 counterpart and redefines the state-of-the-art performance of generative modeling.

REFERENCES

Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22669–22679, June 2023.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions, 2023. URL `https://cdn.openai.com/papers/dall-e-3.pdf`.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22563–22575, June 2023.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=B1xsqj09Fm`.

Andrew Brock, Soham De, and Samuel L Smith. Characterizing signal propagation to close the performance gap in unnormalized resnets. In *International Conference on Learning Representations*, 2021a. URL `https://openreview.net/forum?id=IX3Nnir2omJ`.

Andy Brock, Soham De, Samuel L Smith, and Karen Simonyan. High-performance large-scale image recognition without normalization. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1059–1071. PMLR, 18–24 Jul 2021b. URL `https://proceedings.mlr.press/v139/brock21a.html`.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators, 2024. URL `https://openai.com/research/video-generation-models-as-world-simulators`.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11315–11325, June 2022.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

Francois Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, Rodolphe Jenatton, Lucas Beyer, Michael Tschannen, Anurag Arnab, Xiao Wang, Carlos Riquelme Ruiz, Matthias Minderer, Joan Puigcerver, Utku Evci, Manoj Kumar, Sjoerd Van Steenkiste, Gamaleldin Fathy Elsayed, Aravindh Mahendran, Fisher Yu, Avital Oliver, Fantine Huot, Jasmijn Bastings, Mark Collier, Alexey A. Gritsenko, Vighnesh Birodkar, Cristina Nader Vasconcelos, Yi Tay, Thomas Mensink, Alexander Kolesnikov, Filip Pavetic, Dustin Tran, Thomas Kipf, Mario Lucic, Xiaohua Zhai, Daniel Keysers, Jeremiah J. Harmsen, and Neil Houlsby. Scaling vision transformers to 22 billion parameters. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 7480–7512. PMLR, 23–29 Jul 2023. URL `https://proceedings.mlr.press/v202/dehghani23a.html`.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/esser24a.html.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

Albert Gu, Karan Goel, and Christopher Re. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=uYLFoz1vlAC.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. URL https://openreview.net/forum?id=qw8AKxfYbI.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 8633–8646. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/39235c56aef13fb05a6adc95eb9d8d66-Paper-Conference.pdf.

Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 13213–13232. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/hoogeboom23a.html.

Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017. URL `https://arxiv.org/abs/1704.04861`.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Imagen-Team-Google, :, Jason Baldridge, Jakob Bauer, Mukul Bhutani, Nicole Brichtova, Andrew Bunner, Kelvin Chan, Yichang Chen, Sander Dieleman, Yuqing Du, Zach Eaton-Rosen, Hongliang Fei, Nando de Freitas, Yilin Gao, Evgeny Gladchenko, Sergio Gómez Colmenarejo, Mandy Guo, Alex Haig, Will Hawkins, Hexiang Hu, Huilian Huang, Tobenna Peter Igwe, Christos Kaplanis, Siavash Khodadadeh, Yelin Kim, Ksenia Konyushkova, Karol Langner, Eric Lau, Shixin Luo, Soňa Mokrá, Henna Nandwani, Yasumasa Onoe, Aäron van den Oord, Zarana Parekh, Jordi Pont-Tuset, Hang Qi, Rui Qian, Deepak Ramachandran, Poorva Rane, Abdullah Rashwan, Ali Razavi, Robert Riachi, Hansa Srinivasan, Srivatsan Srinivasan, Robin Strudel, Benigno Uria, Oliver Wang, Su Wang, Austin Waters, Chris Wolff, Auriel Wright, Zhisheng Xiao, Hao Xiong, Keyang Xu, Marc van Zee, Junlin Zhang, Katie Zhang, Wenlei Zhou, Konrad Zolna, Ola Aboubakar, Canfer Akbulut, Oscar Akerlund, Isabela Albuquerque, Nina Anderson, Marco Andreetto, Lora Aroyo, Ben Bariach, David Barker, Sherry Ben, Dana Berman, Courtney Biles, Irina Blok, Pankil Botadra, Jenny Brennan, Karla Brown, John Buckley, Rudy Bunel, Elie Bursztein, Christina Butterfield, Ben Caine, Viral Carpenter, Norman Casagrande, Ming-Wei Chang, Solomon Chang, Shamik Chaudhuri, Tony Chen, John Choi, Dmitry Churbanau, Nathan Clement, Matan Cohen, Forrester Cole, Mikhail Dektiarev, Vincent Du, Praneet Dutta, Tom Eccles, Ndidi Elue, Ashley Feden, Shlomi Fruchter, Frankie Garcia, Roopal Garg, Weina Ge, Ahmed Ghazy, Bryant Gipson, Andrew Goodman, Dawid Górny, Sven Gowal, Khyatti Gupta, Yoni Halpern, Yena Han, Susan Hao, Jamie Hayes, Amir Hertz, Ed Hirst, Tingbo Hou, Heidi Howard, Mohamed Ibrahim, Dirichi Ike-Njoku, Joana Iljazi, Vlad Ionescu, William Isaac, Reena Jana, Gemma Jennings, Donovon Jenson, Xuhui Jia, Kerry Jones, Xiaoen Ju, Ivana Kajic, Christos Kaplanis, Burcu Karagol Ayan, Jacob Kelly, Suraj Kothawade, Christina Kouridi, Ira Ktena, Jolanda Kumakaw, Dana Kurniawan, Dmitry Lagun, Lily Lavitas, Jason Lee, Tao Li, Marco Liang, Maggie Li-Calis, Yuchi Liu, Javier Lopez Alberca, Peggy Lu, Kristian Lum, Yukun Ma, Chase Malik, John Mellor, Inbar Mosseri, Tom Murray, Aida Nematzadeh, Paul Nicholas, João Gabriel Oliveira, Guillermo Ortiz-Jimenez, Michela Paganini, Tom Le Paine, Roni Paiss, Alicia Parrish, Anne Peckham, Vikas Peswani, Igor Petrovski, Tobias Pfaff, Alex Pirozhenko, Ryan Poplin, Utsav Prabhu, Yuan Qi, Matthew Rahtz, Cyrus Rashtchian, Charvi Rastogi, Amit Raul, Ali Razavi, Sylvestre-Alvise Rebuffi, Susanna Ricco, Felix Riedel, Dirk Robinson, Pankaj Rohatgi, Bill Rosgen, Sarah Rumbley, Moonkyung Ryu, Anthony Salgado, Sahil Singla, Florian Schroff, Candice Schumann, Tanmay Shah, Brendan Shillingford, Kaushik Shivakumar, Dennis Shtatnov, Zach Singer, Evgeny Sluzhaev, Valerii Sokolov, Thibault Sottiaux, Florian Stimberg, Brad Stone, David Stutz, Yu-Chuan Su, Eric Tabellion, Shuai Tang, David Tao, Kurt Thomas, Gregory Thornton, Andeep Toor, Cristian Udrescu, Aayush Upadhyay, Cristina Vasconcelos, Alex Vasiloff, Andrey Voynov, Amanda Walker, Luyu Wang, Miaosen Wang, Simon Wang, Stanley Wang, Qifei Wang, Yuxiao Wang, Ágoston Weisz, Olivia Wiles, Chenxia Wu, Xingyu Federico Xu, Andrew Xue, Jianbo Yang, Luo Yu, Mete Yurtoglu, Ali Zand, Han Zhang, Jiageng Zhang, Catherine Zhao, Adilet Zhaxybay, Miao Zhou, Shengqi Zhu, Zhenkai Zhu, Dawn Bloxwich, Mahyar Bordbar, Luis C. Cobo, Eli Collins, Shengyang Dai, Tulsee Doshi, Anca Dragan, Douglas Eck, Demis Hassabis, Sissie Hsiao, Tom Hume, Koray Kavukcuoglu, Helen King, Jack Krawczyk, Yeqing Li, Kathy Meier-Hellstern, Andras Orban, Yury Pinsky, Amar Subramanya, Oriol Vinyals, Ting Yu, and Yori Zwols. Imagen 3, 2024. URL `https://arxiv.org/abs/2408.07009`.

Allan Jabri, David J. Fleet, and Ting Chen. Scalable adaptive computation for iterative generation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 14569–14589. PMLR, 23–29 Jul 2023. URL `https://proceedings.mlr.press/v202/jabri23a.html`.

Alexia Jolicoeur-Martineau, Rémi Piché-Taillefer, Ioannis Mitliagkas, and Remi Tachet des Combes. Adversarial score matching and improved sampling for image generation. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=eLfqMl3z3lq.

Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10124–10134, June 2023.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Hk99zCeAb.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12104–12114. Curran Associates, Inc., 2020a. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/8d30aa96e72440759f74bd2306c1fa3d-Paper.pdf.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020b.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 852–863. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/076ccd93ad68be51f23707988e934906-Paper.pdf.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 26565–26577. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/a98846e9d9cc01cfb87eb694d946ce6b-Paper-Conference.pdf.

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself, 2024a. URL https://arxiv.org/abs/2406.02507.

Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24174–24184, June 2024b.

Diederik Kingma and Ruiqi Gao. Understanding diffusion objectives as the elbo with simple data augmentation. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 65484–65516. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/ce79fbf9baef726645bc2337abb0ade2-Paper-Conference.pdf.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *The Third International Conference on Learning Representations*, 2015.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *The Second International Conference on Learning Representations*, 2014. URL https://openreview.net/forum?id=33X9fd2-9FyZd.

Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/ddeebdeefdb7e7e7a697e1c3e3d8ef54-Paper.pdf.

Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models, 2024. URL https://arxiv.org/abs/2404.07724.

Xiuyu Li, Yijiang Liu, Long Lian, Huanrui Yang, Zhen Dong, Daniel Kang, Shanghang Zhang, and Kurt Keutzer. Q-diffusion: Quantizing diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17535–17545, October 2023a.

Yanyu Li, Huan Wang, Qing Jin, Ju Hu, Pavlo Chemerys, Yun Fu, Yanzhi Wang, Sergey Tulyakov, and Jian Ren. Snapfusion: Text-to-image diffusion model on mobile devices within two seconds. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 20662–20678. Curran Associates, Inc., 2023b. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/41bcc9d3bddd9c90e1f44b29e26d97ff-Paper-Conference.pdf.

Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. In *International Conference on Learning Representations*, 2022a. URL https://openreview.net/forum?id=PlKWVd2yBkY.

Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11976–11986, June 2022b.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan LI, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 5775–5787. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/260a14acce2a89dad36adc8eefe7c59e-Paper-Conference.pdf.

Weijian Luo, Tianyang Hu, Shifeng Zhang, Jiacheng Sun, Zhenguo Li, and Zhihua Zhang. Diff-instruct: A universal approach for transferring knowledge from pre-trained diffusion models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 76525–76546. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/f115f619b62833aadc5acb058975b0e6-Paper-Conference.pdf.

Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14297–14306, June 2023.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed precision training. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1gs9JgRZ.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8162–8171. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/nichol21a.html.

Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16784–16804. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/nichol22a.html.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, October 2023.

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. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=di52zR8xgf.

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=FjNys5c7VyY.

Siyuan Qiao, Huiyu Wang, Chenxi Liu, Wei Shen, and Alan Yuille. Micro-batch training with batch-channel normalization and weight standardization, 2020. URL https://arxiv.org/abs/1903.10520.

Danfeng Qin, Chas Leichner, Manolis Delakis, Marco Fornoni, Shixin Luo, Fan Yang, Weijun Wang, Colby Banbury, Chengxi Ye, Berkin Akin, Vaibhav Aggarwal, Tenghui Zhu, Daniele Moro, and Andrew Howard. Mobilenetv4 – universal models for the mobile ecosystem, 2024. URL https://arxiv.org/abs/2404.10518.

Prajit Ramachandran, Barret Zoph, and Quoc V. Le. Searching for activation functions, 2018. URL https://openreview.net/forum?id=SkBYYyZRZ.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. URL https://arxiv.org/abs/2204.06125.

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 (CVPR)*, pp. 10684–10695, June 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi (eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 36479–36494. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/ec795aeadae0b7d230fa35cbaf04c041-Paper-Conference.pdf.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=TIdIXIpzhoI.

Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P. Kingma. PixelCNN++: Improving the pixelCNN with discretized logistic mixture likelihood and other modifications. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=BJrFC6ceg.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 Conference Proceedings*, SIGGRAPH '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393379. doi: 10.1145/3528233.3530738. URL https://doi.org/10.1145/3528233.3530738.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=St1giarCHLP.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/3001ef257407d5a371a96dcd947c7d93-Paper.pdf.

Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12438–12448. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92c3b916311a5517d9290576e3ea37ad-Paper.pdf.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=PxTIG12RRHS.

George Stein, Jesse Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L Caterini, Eric Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 3732–3784. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/0bc795afae289ed465a65a3b4b1f4eb7-Paper-Conference.pdf.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6105–6114. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/tan19a.html.

Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10096–10106. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/tan21a.html.

Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V. Le. Mnasnet: Platform-aware neural architecture search for mobile. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 7537–7547. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/55053683268957697aa39fba6f231c68-Paper.pdf.

Aaron van den Oord, Nal Kalchbrenner, Lasse Espeholt, koray kavukcuoglu, Oriol Vinyals, and Alex Graves. Conditional image generation with pixelcnn decoders. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/b1301141feffabac455e1f90a7de2054-Paper.pdf.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan LI, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 8406–8441. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/1a87980b9853e84dfb295855b425c262-Paper-Conference.pdf.

Felix Wimbauer, Bichen Wu, Edgar Schoenfeld, Xiaoliang Dai, Ji Hou, Zijian He, Artsiom Sanakoyeu, Peizhao Zhang, Sam Tsai, Jonas Kohler, Christian Rupprecht, Daniel Cremers, Peter Vajda, and Jialiang Wang. Cache me if you can: Accelerating diffusion models through block caching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6211–6220, June 2024.

Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

Jing Nathan Yan, Jiatao Gu, and Alexander M. Rush. Diffusion models without attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8239–8249, June 2024.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6613–6623, June 2024.

Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G. Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, and Lu Jiang. Magvit: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10459–10469, June 2023.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12104–12113, June 2022.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/1e8a19426224ca89e83cef47f1e7f53b-Paper.pdf.

Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Loek7hfb46P.

Barret Zoph and Quoc Le. Neural architecture search with reinforcement learning. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=r1Ue8Hcxg.

Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

A Related Work

We provide a skim-through of several important aspects revolving around diffusion models in prior literature, spanning from training and sampling to network architecture design. We also clarify their similarities and differences compared with our work.

A.1 Diffusion Training

Drawing inspiration from nonequilibrium thermodynamics (Sohl-Dickstein et al., 2015), diffusion models decompose the entire generative process into progressive denoising transitions from standard Gaussian noise to clean images. In stark contrast to other explicit likelihood-based models (*e.g.*, Variational AutoEncoder (Kingma & Welling, 2014), Autoregressive models (van den Oord et al., 2016; Salimans et al., 2017), and Non-Autoregressive models (Chang et al., 2022; Yu et al., 2023)) or implicit likelihood-based models (*e.g.*, GAN (Goodfellow et al., 2014)), diffusion models pose the generation task as a supervised learning scheme, greatly enhancing training stability and thus scalability. In practice, the likelihood-induced Evidence Lower Bound (ELBO) is simplified to an ℓ_2 regression learning objective. Depending on this realization, different regression targets, including image (Sohl-Dickstein et al., 2015), noise (Ho et al., 2020), and velocity (Salimans & Ho, 2022), simply translate to different loss function weights (Kingma & Gao, 2023). In consequence, scaling up diffusion models to billions of parameters and web-scale training data becomes more frictionless compared to the previous prevalent GANs (Kang et al., 2023), incubating a bunch of text-to-image commercial products, such as Stable Diffusion series (Rombach et al., 2022; Podell et al., 2024; Esser et al., 2024), DALL·E 2&3 (Nichol et al., 2022; Ramesh et al., 2022; Betker et al., 2023), and Imagen series (Saharia et al., 2022; Imagen-Team-Google et al., 2024). In the present work, we inherit the EDM (Karras et al., 2022) preconditioning framework for training due to its well-behaved training dynamics.

A.2 Diffusion Sampling

For diffusion models, the sampling procedure typically demands thousands of consecutive steps to synthesize a high-quality image. Theoretically, the diffusion backward process could be interpreted as a reverse Stochastic Differential Equation (SDE) or the corresponding Probability Flow Ordinary Differential Equation (PF-ODE) (Song et al., 2021b). Therefore, there naturally exists a trade-off between the discretion error and step size. The straighter the sampling trajectory, the larger step size can be tolerated. As such, much endeavor has been devoted to straightening the trajectory (Song et al., 2021a; Karras et al., 2022; Liu et al., 2023) and inventing advanced ODE solvers (Liu et al., 2022a; Lu et al., 2022; Zhang & Chen, 2023). In parallel, a growing body of diffusion distillation techniques (Salimans & Ho, 2022; Meng et al., 2023; Luo et al., 2023; Yin et al., 2024) is proposed to reduce the number of sampling steps. In addition, the overall sampling cost could also be reduced by cutting down the model latency per step in the denoising trajectory. Our work goes along this research vein, contributing to a compact yet high-performing model via reworking the network structure from scratch. Hence, our work distinguishes itself from post-hoc pruning (Li et al., 2023b), quantization (Li et al., 2023a) and cache (Wimbauer et al., 2024) methodology but is complementary.

A.3 Network Architecture Engineering

U-Net is ubiquitously applied to low-level vision tasks, including visual segmentation (Chen et al., 2018), generation (Kingma et al., 2016) and so on. Skip connections in the network always play an instrumental role in transmitting the high-resolution signal to the output end for detailed refinement, either in CNN or Transformer (Bao et al., 2023). Building upon a U-Net backbone, DDPM (Ho et al., 2020) interleaves convolution blocks with self-attention modules (Vaswani et al., 2017), effectively gathering long-range pixel dependence. iDDPM (Nichol & Dhariwal, 2021) extends single-head self-attention to multi-head ones and widens its usage over a broader range of feature resolutions. Adaptive Group Norm (AdaGN) is involved as well, resembling AdaIN (Huang & Belongie, 2017). ADM (Dhariwal & Nichol, 2021) additionally steals the network topology and scaled residual connections from the GAN literature (Brock et al., 2019; Karras et al., 2020b). Several hyperparameters are ablated here, including the network depth/width and the number of attention heads. EDM2 (Karras et al., 2024b) emphasizes standardizing the magnitudes of network weights, activations, gradients, *etc*., in the same spirit of pioneering magnitude-focusing image recognition networks (Brock et al., 2021a;b). Diffusion Transformers (DiT) position themselves as appealing alternatives to the *de facto* standard U-Net, attracting enthusiasm from both academia and industry (Peebles & Xie, 2023; Hoogeboom et al., 2023). Subsequently, DiffuSSM (Yan et al., 2024) supplants the attention mechanism of DiT with the State Space Model (SSM) (Gu et al., 2022) blocks to promote the efficiency. Our work takes root in a hybrid architecture, EDM2, that alternates convolution blocks with attention modules and manifests as an outstanding denoiser architecture.