EFFICIENT AND SCALABLE GENERATIVE MODEL CONTROL FOR HIGH-QUALITY MULTIMODAL SYNTHESIS

by
Kangfu Mei

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
01 2025

Abstract

Generative models, such as diffusion models and generative adversarial networks, have recently transformed foundational vision tasks, including generating images from noise. However, challenges remain in designing generative models that are best suited for real-world applications and in improving their generalization capabilities across downstream tasks. This thesis addresses these challenges through comprehensive theoretical analyses and empirical experiments, with a focus on practical visual perception rectification tasks.

First, the thesis investigates the scaling properties of latent diffusion models, widely used in text-to-image generation and its downstream applications. It introduces inference scaling laws that reveal the surprising superiority of small models over large models when leveraging increased inference compute. Next, it presents a novel scalable video diffusion model capable of generating continuous scenes from pure noise, extending generative capabilities from static images to dynamic, expressive videos. To further reduce the complexity of designing modality-specific models, the thesis proposes a versatile field diffusion model that can seamlessly handle various modalities, including image, video, 3D, and game environments. Additionally, the thesis introduces an efficient diffusion distillation technique that achieves comparable visual quality while reducing computational cost by 99%, significantly enhancing the sampling efficiency of generative models.

Building on these advancements, the thesis applies realism priors derived from genera-

tive models to three real-world image processing tasks: turbulence removal, shadow removal, and single-image super-resolution. These approaches consistently achieve state-of-the-art performance, surpassing traditional regression-based methods and producing results with enhanced realism. This work sets a new benchmark for rectifying visual content while preserving natural details. Finally, the thesis explores future directions for advancing generative models toward the ultimate goal of simulating and controlling virtual worlds.

Keywords: Diffusion Models, GAN Inversion, Scaling Laws, Diffusion Distillation, Multi-modal Vision-Language Model, Video Generation, Image Processing

Primary reader and thesis advisor

Dr. Vishal M. Patel
Professor
Department of Electrical and Computer Engineering
Johns Hopkins University, Baltimore MD

Secondary readers

Dr. Rama Chellappa
Professor
Department of Electrical and Computer Engineering
Johns Hopkins University, Baltimore, MD

Dr. Alan Yuille
Professor
Department of Computer Science
Johns Hopkins University, Baltimore, MD

This thesis is dedicated to my grandfather who told me I should go into doctor degrees. This wasn't what we were expecting, but it's what worked!

Acknowledgement

First and foremost, I extend my deepest gratitude to my PhD advisor, Prof. Vishal M. Patel, for graciously accepting me as his PhD student at Johns Hopkins University. His unwavering inspiration and encouragement have been pivotal in shaping my academic journey and nurturing my aspirations.

I am equally grateful to my committee members, Prof. Rama Chellappa and Prof. Alan Yuille, whose extraordinary contributions to the field have profoundly influenced and guided my own research endeavors.

I also wish to extend my sincere thanks to Mauricio Delbracio, Hossein Talebi, and Peyman Milanfar for hosting me at Google Research in Mountain View during 2023 and 2024. Collaborating on the paper exploring diffusion distillation, diffusion scaling properties, and multimodal super-resolution was a true privilege.

Additionally, my heartfelt appreciation goes to my family, collaborators, labmates, and friends, whose unwavering support has made this journey possible.

Lastly, I want to acknowledge my younger self for choosing to embark on this challenging yet rewarding life experience.

Table of Contents

List of Tables

List of Figures

Chapter 1
Introduction and Background

1.1 Diffusion Models

Diffusion Models. A diffusion model [1, 2] has latent variables $\{\mathbf{z}_t | t \in [0, T]\}$ specified by a noise schedule comprising differentiable functions $\{\alpha_t, \sigma_t\}$ with $\sigma_t^2 = 1 - \alpha_t^2$. The clean data $\mathbf{x} \sim p_{\text{data}}$ is progressively perturbed in a (forward) Gaussian process as in the following Markovian structure:

$$q(\mathbf{z}_t|\mathbf{x}) = \mathcal{N}(\mathbf{z}_t; \alpha_t \mathbf{x}, \sigma_t^2 \mathbf{I}), \tag{1.1}$$

$$q(\mathbf{z}_t|\mathbf{z}_s) = \mathcal{N}(\mathbf{z}_t; \alpha_{t|s} \mathbf{z}_s, \sigma_{t|s}^2 \mathbf{I}), \tag{1.2}$$

where $0 \leq s < t \leq 1$ and $\alpha_{t|s}^2 = \alpha_t / \alpha_s$. Here the latent $\mathbf{z}_t$ is sampled from the combination of the clean data and random noise by using the reparameterization trick [3], which has $\mathbf{z}_t = \alpha_t \mathbf{x} + \sigma_t \epsilon$.

Deterministic sampling. The aforementioned diffusion process that starts from $\mathbf{z}_0 \sim p_{\text{data}}(\mathbf{x})$ and ends at $\mathbf{z}_T \sim \mathcal{N}(0, \mathbf{I})$ can be modeled as the solution of an stochastic differential equation (SDE) [1]. The SDE is formed by a vector-value function $f(\cdot, \cdot) : \mathbb{R}^d \to \mathbb{R}^d$, a scalar function $g(\cdot) : \mathbb{R} \to \mathbb{R}$, and the standard Wiener process $\mathbf{w}$ as:

$$\mathrm{d}\mathbf{z}_t = f(\mathbf{z}_t, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}. \tag{1.3}$$

The overall idea is that the reverse-time SDE that runs backwards in time, can generate samples of p_{data} from the prior distribution $\mathcal{N}(0, \mathbf{I})$. This reverse SDE is given by

$$\mathrm{d}\mathbf{z}_t = [f(\mathbf{z}_t, t) - g(t)^2 \nabla_{\mathbf{z}} \log p_t(\mathbf{z}_t)]\mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}, \tag{1.4}$$

where the $\bar{\mathbf{w}}$ is a also standard Wiener process in reversed time, and $\nabla_{\mathbf{z}} \log p_t(\mathbf{z}_t)$ is the score of the marginal distribution at time t. The score function can be estimated by training a score-based model $s_\theta(\mathbf{z}_t, t) \approx \nabla_z \log p_t(\mathbf{z}_t)$ with score-matching [4] or a denoising network $\hat{\mathbf{x}}_\theta(\mathbf{z}_t, t)$ [2]:

$$s_\theta(\mathbf{z}_t, t) := (\alpha_t \hat{\mathbf{x}}_\theta(\mathbf{z}_t, t) - \mathbf{z}_t)/\sigma_t^2. \tag{1.5}$$

Such backward SDE satisfies a special ordinary differential equation (ODE) that allows deterministic sampling given $\mathbf{z}_T \sim \mathcal{N}(0, \mathbf{I})$. This is known as the *probability flow* (PF) ODE [1] and is given by

$$\mathrm{d}\mathbf{z}_t = [f(\mathbf{z}_t, t) - \frac{1}{2} g^2(t) s_\theta(\mathbf{z}_t, t)]\mathrm{d}t, \tag{1.6}$$

where $f(\mathbf{z}_t, t) = \frac{\mathrm{d}\log\alpha_t}{\mathrm{d}t}\mathbf{z}_t$, $g^2(t) = \frac{\mathrm{d}\sigma_t^2}{\mathrm{d}t} - 2\frac{\mathrm{d}\log\alpha_t}{\mathrm{d}t}\sigma_t^2$ with respect to $\{\alpha_t, \sigma_t\}$ and t according to [5]. This ODE can be solved numerically with diffusion samplers like DDIM [6], where starting from $\hat{\mathbf{z}}_T \sim \mathcal{N}(0, \mathbf{I})$, we update for $s = t - \Delta t$:

$$\hat{\mathbf{z}}_s := \alpha_s \hat{\mathbf{x}}_\theta(\hat{\mathbf{z}}_t, t) + \sigma_s(\hat{\mathbf{z}}_t - \alpha_t \hat{\mathbf{x}}_\theta(\hat{\mathbf{z}}_t, t))/\sigma_t, \tag{1.7}$$

till we reach $\hat{\mathbf{z}}_0$.

Diffusion models parametrizations. Leaving aside the aforementioned way of parametrizing diffusion models with a denoising network (signal prediction) or a score model (noise prediction equation 1.5), in this work, we adopt a parameterization that mixes both the score (or noise) and the signal prediction. Existing methods include either predicting the noise $\hat{\epsilon}_\theta(\mathbf{x}_t, t)$ and the signal $\hat{\mathbf{x}}_\theta(\mathbf{z}_t, t)$ separately using a single network [7], or predicting a combination of noise and signal by expressing them in a new term, like the velocity model $\hat{\mathbf{v}}_\theta(\mathbf{z}_t, t) \approx \alpha_t \epsilon - \sigma_t \mathbf{x}$ [8]. Note that one can derive an estimation of the signal and the noise from the velocity one,

$$\hat{\mathbf{x}} = \alpha_t \mathbf{z}_t - \sigma_t \hat{\mathbf{v}}_\theta(\mathbf{z}_t, t), \text{ and } \hat{\epsilon} = \alpha_t \hat{\mathbf{v}}_\theta(\mathbf{z}_t, t) + \sigma_t \mathbf{z}_t. \tag{1.8}$$

Similarly, DDIM update rule (equation 1.7) can be rewritten in terms of the velocity parametrization:

$$\hat{\mathbf{z}}_s := \alpha_s(\alpha_t \hat{\mathbf{z}}_t - \sigma_t \hat{\mathbf{v}}_\theta(\hat{\mathbf{z}}_t, t)) + \sigma_s(\alpha_t \hat{\mathbf{v}}_\theta(\hat{\mathbf{z}}_t, t) + \sigma_t \hat{\mathbf{z}}_t). \tag{1.9}$$

Chapter 2

Scaling Properties of Latent Diffusion Models

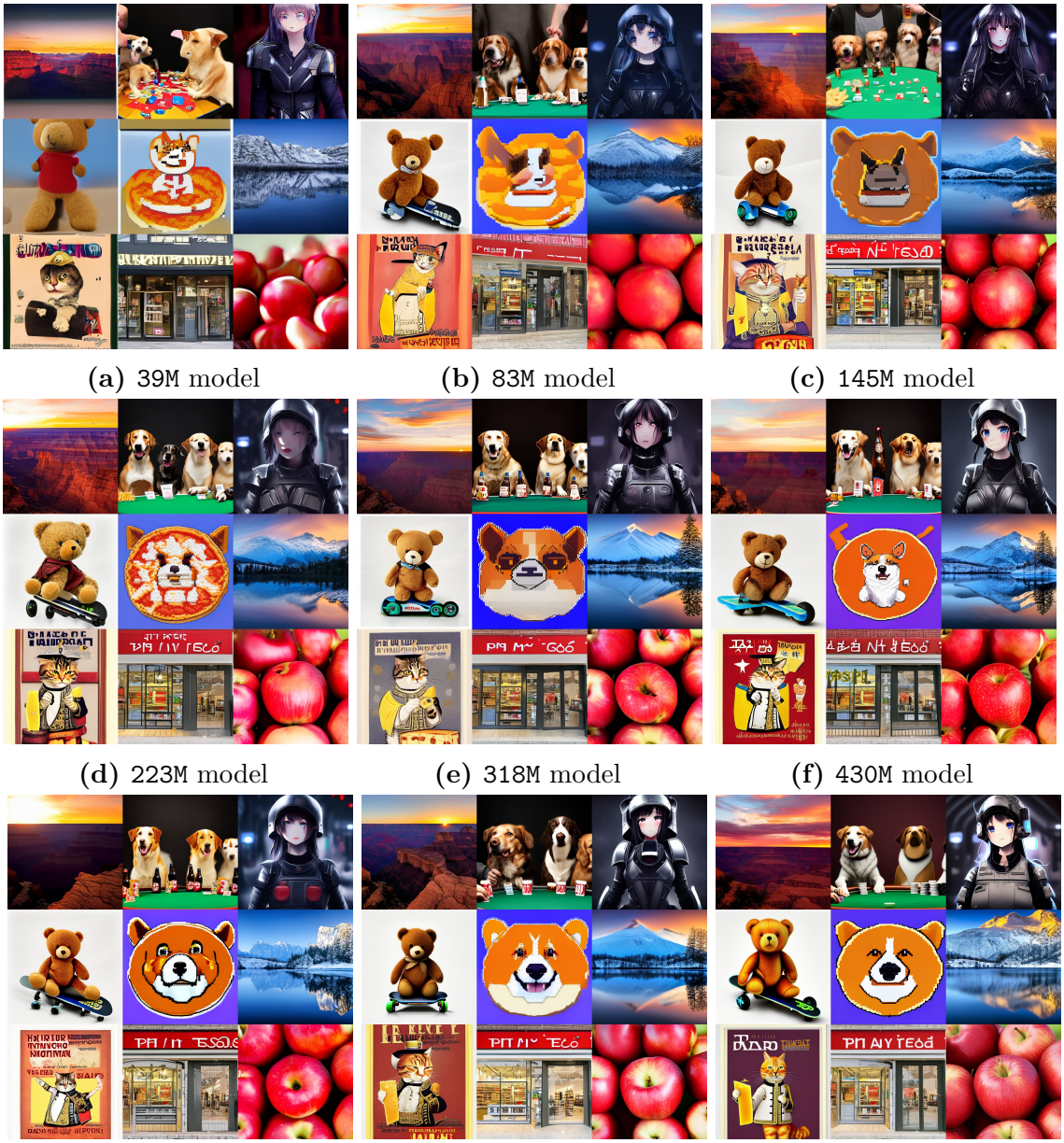
2.1 Introduction

Latent diffusion models (LDMs) [9], and diffusion models in general, trained on large-scale, high-quality data [10, 11] have emerged as a powerful and robust framework for generating impressive results in a variety of tasks, including image synthesis and editing [9, 12–15], video creation [16–19], audio production [20], and 3D synthesis [21, 22]. Despite their versatility, the major barrier against wide deployment in real-world applications [23, 24] comes from their low *sampling efficiency*. The essence of this challenge lies in the inherent reliance of LDMs on multi-step sampling [1, 2] to produce high-quality outputs, where the total cost of sampling is the product of sampling steps and the cost of each step. Specifically, the go-to approach involves using the 50-step DDIM sampling [6, 9], a process that, despite ensuring output quality, still requires a relatively long latency for completion on modern mobile devices with post-quantization. In contrast to single shot generative models (e.g., generative-adversarial networks (GANs) [25]) which bypass the need for iterative refinement [25, 26], the operational latency of LDMs calls for a pressing need for efficiency optimization to further facilitate their practical applications.

Recent advancements in this field [24, 27–31] have primarily focused on developing faster network architectures with comparable model size to reduce the inference time per step, along with innovations in improving sampling algorithms that allow

for using less sampling steps [6, 32–36]. Further progress has been made through diffusion-distillation techniques [8, 37–41], which simplifies the process by learning multi-step sampling results in a single forward pass, and then broadcasts this single-step prediction multiple times. These distillation techniques leverage the redundant learning capability in LDMs, enabling the distilled models to assimilate additional distillation knowledge. Despite these efforts being made to improve diffusion models, the sampling efficiency of smaller, less redundant models has not received adequate attention. A significant barrier to this area of research is the scarcity of available modern accelerator clusters [42], as training high-quality text-to-image (T2I) LDMs from scratch is both time-consuming and expensive—often requiring several weeks and hundreds of thousands of dollars.

In this chapter, we empirically investigate the scaling properties of LDMs, with a particular focus on understanding how their scaling properties impact the sampling efficiency across various model sizes. We trained a suite of 12 text-to-image LDMs from scratch, ranging from 39 million to 5 billion parameters, under a constrained budget. Example results are depicted in Fig. 2.1. All models were trained on TPUv5 using internal data sources with about 600 million aesthetically-filtered text-to-image pairs. Our study reveals that there exist a scaling trend within LDMs, notably that smaller models may have the capability to surpass larger models under an equivalent sampling budget. Furthermore, we investigate how the size of pre-trained text-to-image LDMs affects their sampling efficiency across diverse downstream tasks, such as real-world super-resolution [43, 44] and subject-driven text-to-image synthesis (i.e., Dreambooth) [45].



(a) 39M model **(b)** 83M model **(c)** 145M model

(d) 223M model **(e)** 318M model **(f)** 430M model

(g) 558M model **(h)** 704M model **(i)** 2B model

Figure 2.1: Text-to-image results from our scaled LDMs (39M - 2B), highlighting the improvement in visual quality with increased model size (note: 39M model is the exception). All images generated using 50-step DDIM sampling and CFG rate of 7.5.

2.1.1 Summary

Our key findings for scaling latent diffusion models in text-to-image generation and various downstream tasks are as follows:

Pretraining performance scales with training compute. We demonstrate a clear link between compute resources and LDM performance by scaling models from 39 million to 5 billion parameters. This suggests potential for further improvement with increased scaling. See Section 2.2.1 for details.

Downstream performance scales with pretraining. We demonstrate a strong correlation between pretraining performance and success in downstream tasks. Smaller models, even with extra training, cannot fully bridge the gap created by the pretraining quality of larger models. This is explored in detail in Section 2.2.2.

Smaller models sample more efficient. Smaller models initially outperform larger models in image quality for a given sampling budget, but larger models surpass them in detail generation when computational constraints are relaxed. This is further elaborated in Section 2.2.3.

Sampler does not change the scaling efficiency. Smaller models consistently demonstrate superior sampling efficiency, regardless of the diffusion sampler used. This holds true for deterministic DDIM [6], stochastic DDPM [2], and higher-order DPM-Solver++ [46]. For more details, see Section 2.2.3.

Smaller models sample more efficient on the downstream tasks with fewer steps. The advantage of smaller models in terms of sampling efficiency extends to the

downstream tasks when using less than 20 sampling steps. This is further elaborated in Section 2.2.4.

Diffusion distillation does not change scaling trends. Even with diffusion distillation, smaller models maintain competitive performance against larger distilled models when sampling budgets are constrained. This suggests distillation does not fundamentally alter scaling trends. See Section 2.2.5 for in-depth analysis.

2.2 Scaling LDMs

We developed a family of powerful Latent Diffusion Models (LDMs) built upon the widely-used `866M` Stable Diffusion v1.5 standard [9]. The denoising UNet of our models offers a flexible range of sizes, with parameters spanning from `39M` to `5B`. We incrementally increase the number of filters in the residual blocks while maintaining other architecture elements the same, enabling a predictably controlled scaling. Table 2.1 shows the architectural differences among our scaled models. We also provide the relative cost of each model against the baseline model. Fig. 2.2 shows the architectural differences during scaling. Models were trained using the web-scale aesthetically filtered text-to-image dataset, *i.e.*, WebLI [47]. All the models are trained for 500K steps, batch size 2048, and learning rate 1e-4. This allows for all the models to have reached a point where we observe diminishing returns. Fig. 2.1 demonstrates the consistent generation capabilities across our scaled models. We used the common practice of 50 sampling steps with the DDIM sampler, 7.5 classifier-free guidance rate, for text-to-image generation. The visual quality of the results exhibits a clear improvement as model size increases.

Params	39M	83M	145M	223M	318M	430M	558M	704M	866M	2B	5B
Filters (c)	64	96	128	160	192	224	256	288	320	512	768
GFLOPS	25.3	102.7	161.5	233.5	318.5	416.6	527.8	652.0	789.3	1887.5	4082.6
Norm. Cost	0.07	0.13	0.20	0.30	0.40	0.53	0.67	0.83	1.00	2.39	5.17
FID ↓	25.30	24.30	24.18	23.76	22.83	22.35	22.15	21.82	21.55	20.98	20.14
CLIP ↑	0.305	0.308	0.310	0.310	0.311	0.312	0.312	0.312	0.312	0.312	0.314

Table 2.1: We scale the baseline LDM (*i.e.*, 866M Stable Diffusion v1.5) by changing the base number of channels c that controls the rest of the U-Net architecture as $[c, 2c, 4c, 4c]$ (See Fig. 2.2). GFLOPS are measured for an input latent of shape $64 \times 64 \times 4$ with FP32. We also show a normalized running cost with respect to the baseline model. The text-to-image performance (FID and CLIP scores) for all scaled LDMs is evaluated on the COCO-2014 validation set with 30k samples, using 50-step DDIM sampling and Classifier-free Guidance (CFG) with a rate of 7.5. It is worth noting that all the model sizes, and the training and the inference costs reported in this work only refer to the denoising UNet in the latent space, and do not include the 1.4B text encoder and the 250M latent encoder and decoder.

In order to evaluate the performance of the scaled models, we test the text-to-image performance of scaled models on the validation set of COCO 2014 [10] with 30k samples. For downstream performance, specifically real-world super-resolution, we test the performance of scaled models on the validation of DIV2K with 3k randomly cropped patches, which are degraded with the RealESRGAN degradation [48].

2.2.1 Training compute scales text-to-image performance

We find that our scaled LDMs, across various model sizes, exhibit similar trends in generative performance relative to training compute cost, especially after training stabilizes, which typically occurs after 200K iterations. These trends demonstrate a smooth scaling in learning capability between different model sizes. To elaborate,

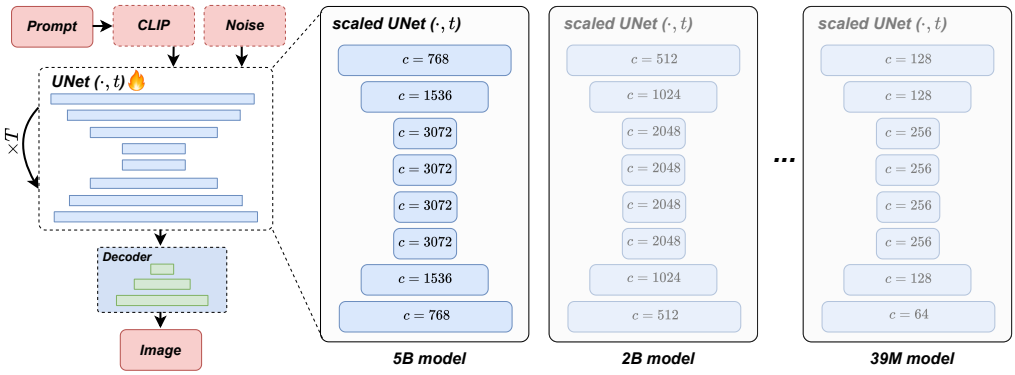


Figure 2.2: Our scaled latent diffusion models vary in the number of filters within the denoising U-Net. Other modules remain consistent. Smooth channel scaling (64 to 768) within residual blocks yields models ranging from `39M` to `5B` parameters. For downstream tasks requiring image input, we use an encoder to generate a latent code; this code is then concatenated with the noise vector in the denoising U-Net.

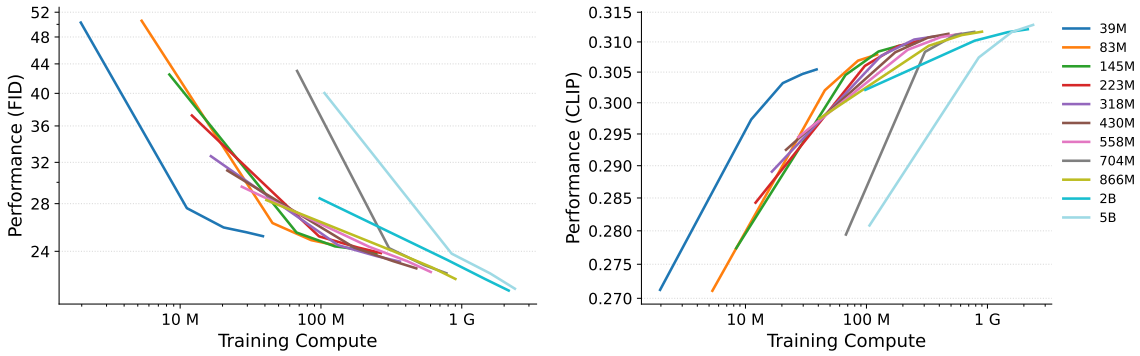


Figure 2.3: In text-to-image generation using 50-step DDIM sampling and CFG rate of 7.5, we observe consistent trends across various model sizes in how quality metrics (FID and CLIP scores) relate to training compute (*i.e.*, the total GFLOPS spend on training). Under moderate training resources, training compute is the most relevant factor dominating quality.

Fig. 2.3 illustrates a series of training runs with models varying in size from 39 million to 5 billion parameters, where the training compute cost is quantified as the product of relative cost shown in Table 2.1 and training iterations. Model performance is evaluated by using the same sampling steps and sampling parameters. In scenarios

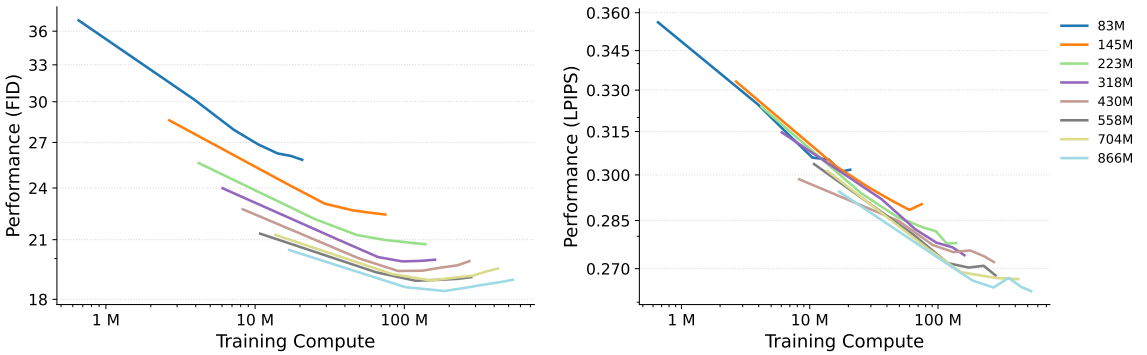


Figure 2.4: In 4× real image super-resolution using 50-step DDIM sampling, FID and LPIPS scores reveal an interesting divergence. Model size drives FID score improvement, while training compute most impacts LPIPS score. Despite this, visual assessment (Fig. 2.5) confirms the importance of model size for superior detail recovery (similarly as observed in the text-to-image pretraining).

with moderate training compute (i.e., $< 1G$, see Fig. 2.3), the generative performance of T2I models scales well with additional compute resources.

2.2.2 Pretraining scales downstream performance

Using scaled models based on their pretraining on text-to-image data, we finetune these models on the downstream tasks of real-world super-resolution [43, 44] and DreamBooth [45]. The performance of these pretrained models is shown in Table. 2.1. In the left panel of Fig. 2.4, we present the generative performance FID versus training compute on the super-resolution (SR) task. It can be seen that the performance of SR models is more dependent on the model size than training compute. Our results demonstrate a clear limitation of smaller models: they cannot reach the same performance levels as larger models, regardless of training compute.

While the distortion metric LPIPS shows some inconsistencies compared to the

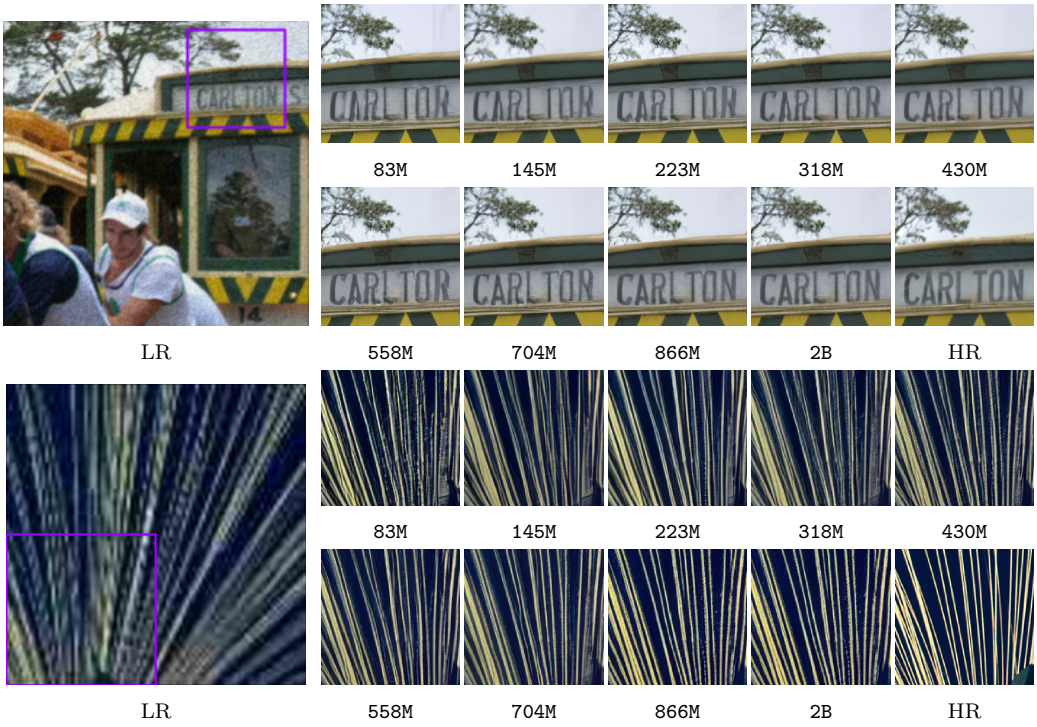


Figure 2.5: In 4× super-resolution using 50-step DDIM sampling, visual quality directly improves with increased model size. As these scaled models vary in pretraining performance, the results clearly demonstrate that pretraining boosts super-resolution capabilities in both quantitative (Fig 2.4) and qualitative ways.

generative metric FID (Fig. 2.4), Fig. 2.5 clearly demonstrates that larger models excel in recovering fine-grained details compared to smaller models.

The key takeaway from Fig. 2.4 is that large super-resolution models achieve superior results even after short finetuning periods compared to smaller models. This suggests that pretraining performance (dominated by the pretraining model sizes) has a greater influence on the super-resolution FID scores than the duration of finetuning (*i.e.*, training compute for finetuning). Furthermore, we compare the visual results of the DreamBooth finetuning on the different models in Fig. 2.6. We observe a similar trend

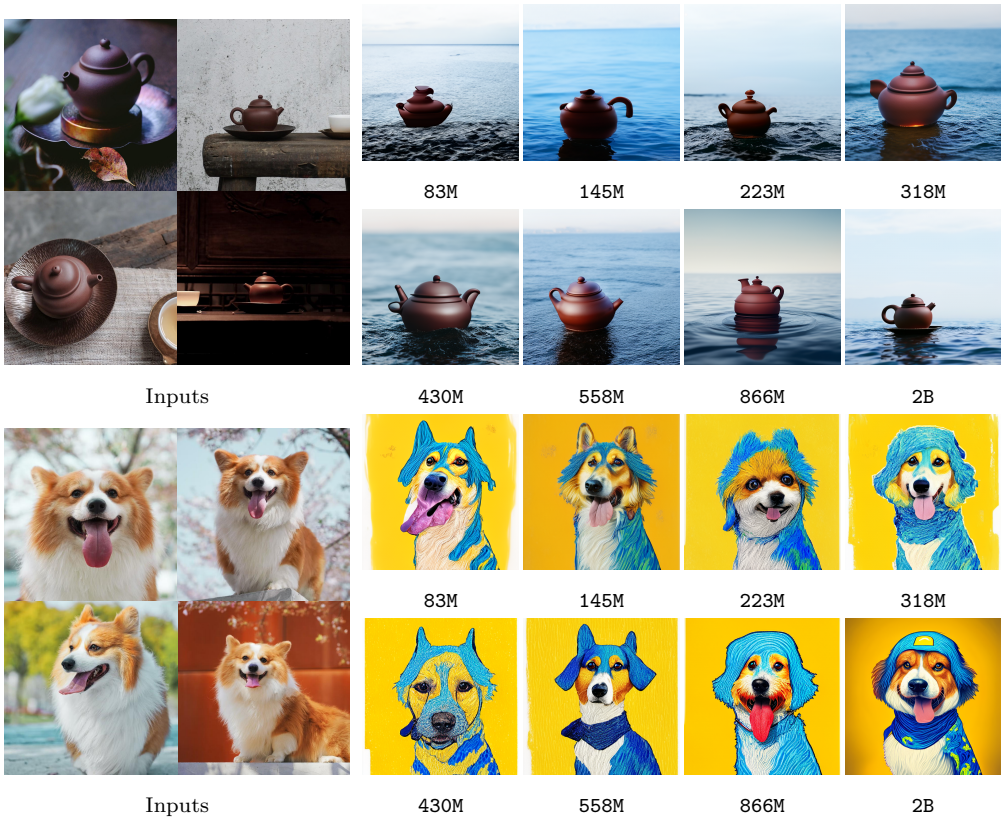


Figure 2.6: Visualization of the Dreambooth results (using 50-step DDIM sampling and CFG rate of 7.5) shows two distinct tiers based on model size. Smaller models (`83M-223M`) perform similarly, as do larger ones (`318M-2B`), with a clear quality advantage for the larger group.

between visual quality and model size.

2.2.3 Scaling sampling-efficiency

Analyzing the effect of CFG rate. Text-to-image generative models require nuanced evaluation beyond single metrics. Sampling parameters are vital for customization, with the Classifier-Free Guidance (CFG) rate [49] directly influencing the balance between visual fidelity and semantic alignment with text prompt. Rombach

(a) 50-step sampling results of the 145M model

(b) 50-step sampling results of the 866M model

Figure 2.7: Visualization of text-to-image results with 50-step DDIM sampling and different CFG rates (from left to right in each row: $(1.5, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0)$). The prompt used is "*A raccoon wearing formal clothes, wearing a top hat and holding a cane. Oil painting in the style of Rembrandt.*". We observe that changes in CFG rates impact visual quality more significantly than the prompt semantic accuracy. We use the FID score for quantitative determination of optimal sampling performance (Fig. 2.8) because it directly measures visual quality, unlike the CLIP score, which focuses on semantic similarity.

et al. [9] experimentally demonstrate that different CFG rates result in different CLIP and FID scores.

In this study, we find that CFG rate as a sampling parameter yields inconsistent results across different model sizes. Hence, it is interesting to quantitatively determine the *optimal* CFG rate for each model size and sampling steps using either FID or CLIP score. We demonstrate this by sampling the scaled models using different CFG rates, *i.e.*, $(1.5, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0)$ and comparing their quantitative and qualitative results. In Fig. 2.7, we present visual results of two models under varying CFG rates, highlighting the impact on the visual quality. We observed that changes in CFG rates impact visual quality more significantly than prompt semantic accuracy and therefore opted to use the FID score for quantitative determination of the optimal CFG rate.

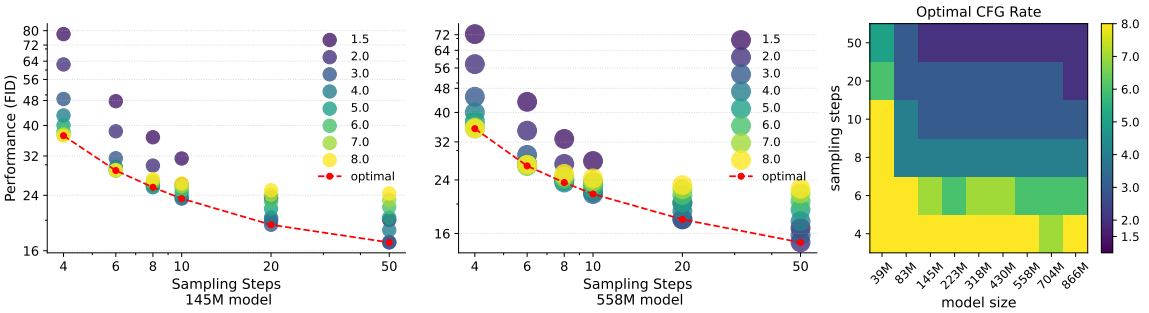


Figure 2.8: The impact of the CFG rate on text-to-image generation depends on the model size and sampling steps. As demonstrated in the left and center panels, the optimal CFG rate changes as the sampling steps increased. To determine the optimal performance (according to the FID score) of each model and each sampling steps, we systematically sample the model at various CFG rates and identify the best one. As a reference of the optimal performance, the right panel shows the CFG rate corresponding to the optimal performance of each model for a given number of sampling steps.

performance. Fig. 2.8 shows how different classifier-free guidance rates affect the FID scores in text-to-image generation (see figure caption for more details).

Scaling efficiency trends. Using the optimal CFG rates established for each model at various number of sampling steps, we analyze the optimal performance to understand the sampling efficiency of different LDM sizes. Specifically, in Fig. 2.9, we present a comparison between different models and their optimal performance given the sampling cost (normalized cost $\times$ sampling steps). By tracing the points of optimal performance across various sampling cost—represented by the dashed vertical line—we observe a consistent trend: smaller models frequently outperform larger models across a range of sampling cost in terms of FID scores. Furthermore, to visually substantiate better-quality results generated by smaller models against larger ones, Fig. 2.10 compares the results of different scaled models, which highlights that

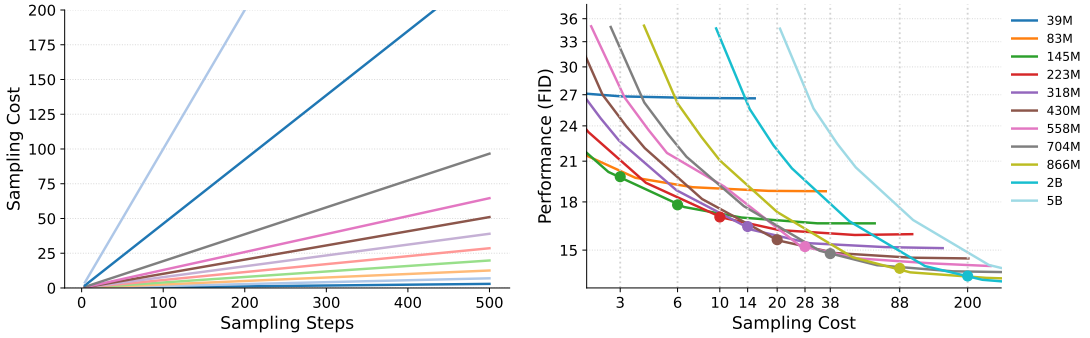


Figure 2.9: Comparison of text-to-image performance of models with varying sizes. The left figure shows the relationship between sampling cost (normalized cost $\times$ sampling steps) and sampling steps for different model sizes. The right figure plots the optimal text-to-image FID score among CFG rates of $(1.5, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0)$ as a function of the sampling cost for the same models. Key Observation: Smaller models achieve better FID scores than larger models for a fixed sampling cost. For instance, at a cost of 3, the 83M model achieves the best FID compared to the larger models. This suggests that smaller models can be more efficient in achieving good results with lower costs.

the performance of smaller models can indeed match their larger counterparts under similar sampling cost conditions.

Scaling sampling-efficiency in different samplers To assess the generalizability of observed scaling trends in sampling efficiency, we compared scaled LDM performance using different diffusion samplers. In addition to the default DDIM sampler, we employed two representative alternatives: the stochastic DDPM sampler [2] and the high-order DPM-Solver++ [46].

Experiments illustrated in Fig. 2.11 reveal that the DDPM sampler typically produces lower-quality results than DDIM with fewer sampling steps, while the DPM-Solver++ sampler generally outperforms DDIM in image quality (see the figure caption for details). Importantly, we observe consistent sampling-efficiency trends with the DDPM

(a) Prompt: *"A corgi's head depicted as a nebula."*. Sampling Cost ≈ 6.

(b) Prompt: *"A pineapple surfing on a wave."*. Sampling Cost ≈ 12.

Figure 2.10: Text-to-image results of the scaled LDMs under approximately the same inference cost (normalized cost $\times$ sampling steps). Smaller models can produce comparable or even better visual results than larger models under similar sampling cost.

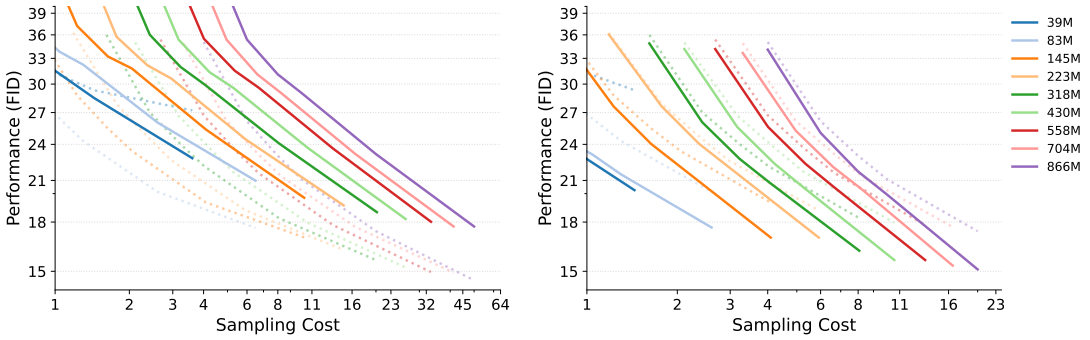


Figure 2.11: *Left*: Text-to-image performance FID as a function of the sampling cost (normalized cost × sampling steps) for the DDPM sampler (solid curves) and the DDIM sampler (dashed curves). *Right*: Text-to-image performance FID as a function of the sampling cost for the second-order DPM-Solver++ sampler (solid curves) and the DDIM sampler (dashed curves). Suggested by the trends shown in Fig. 2.9, we only show the sampling steps ≤ 50 as using more steps does not improve the performance.

and DPM-Solver++ sampler as seen with the default DDIM: smaller models tend to achieve better performance than larger models under the same sampling cost. Since the DPM-Solver++ sampler is not designed for use beyond 20 steps, we focused our testing within this range. This finding demonstrates that the scaling properties of LDMs remain consistent regardless of the diffusion sampler used.

2.2.4 Scaling downstream sampling-efficiency

Here, we investigate the scaling sampling-efficiency of LDMs on downstream tasks, specifically focusing on the super-resolution task. Unlike our earlier discussions on optimal sampling performance, there is limited literature demonstrating the positive impacts of SR performance without using classifier-free guidance. Thus, our approach directly uses the SR sampling result without applying classifier-free guidance. Inspired from Fig. 2.4, where the scaled downstream LDMs have significant performance

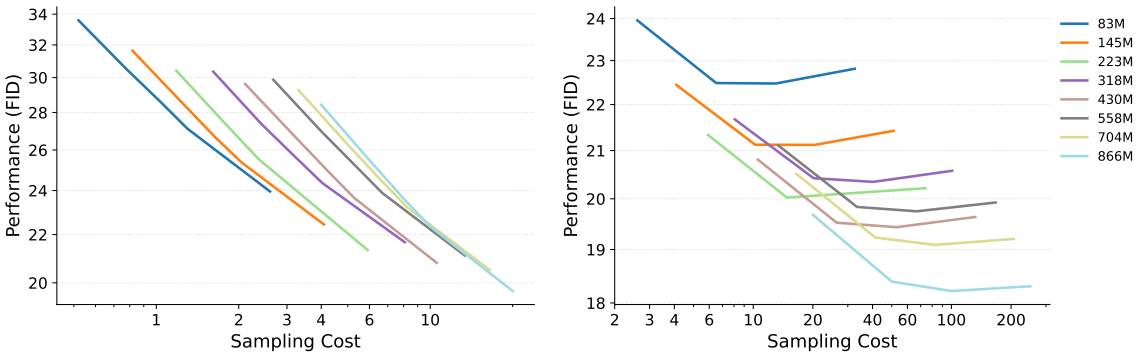


Figure 2.12: Super-resolution performance vs. sampling cost for different model sizes. *Left:* FID scores of super-resolution models under limited sampling steps (less than or equal to 20). Smaller models tend to achieve lower (better) FID scores within this range. *Right:* FID scores of super-resolution models under a larger number of sampling steps (greater than 20). Performance differences between models become less pronounced as sampling steps increase.

difference in 50-step sampling, we investigate sampling efficiency from two different aspects, *i.e.*, fewer sampling steps $[4, 20]$ and more sampling steps $(20, 250]$. As shown in the left part of Fig. 2.12, the scaling sampling-efficiency still holds in the SR tasks when the number of sampling steps is less than or equal to 20 steps. Beyond this threshold, however, larger models demonstrate greater sampling-efficiency than smaller models, as illustrated in the right part of Fig. 2.12. This observation suggests the consistent sampling efficiency of scaled models on fewer sampling steps from text-to-image generation to super-resolution tasks.

2.2.5 Scaling sampling-efficiency in distilled LDMs.

We have featured the scaling sampling-efficiency of latent diffusion models, which demonstrates that smaller model sizes exhibit higher sampling efficiency. A notable caveat, however, is that smaller models typically imply reduced modeling capability.

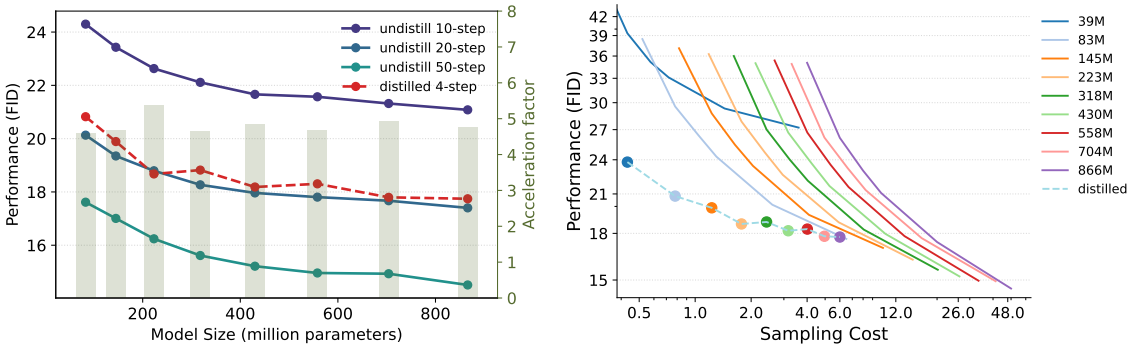


Figure 2.13: Distillation improves text-to-image performance and scalability. *Left:* Distilled Latent Diffusion Models (LDMs) consistently exhibit lower (better) FID scores compared to their undistilled counterparts across varying model sizes. The consistent acceleration factor (approx. 5×) indicates that the benefits of distillation scale well with model size. *Right:* Distilled models using only 4 sampling steps achieve FID scores comparable to undistilled models using significantly more steps. Interestingly, at a sampling cost of 7, the distilled `866M` model performs similarly to the smaller, undistilled `83M` model, suggesting improved efficiency.

This poses a challenge for recent diffusion distillation methods [8, 37–41, 50, 51] that heavily depend on modeling capability. One might expect a contradictory conclusion and believe the distilled large models sample faster than distilled small models. In order to demonstrate the sampling efficiency of scaled models after distillation, we distill our previously scaled models with conditional consistency distillation [38, 41] on text-to-image data and compare those distilled models on their optimal performance.

To elaborate, we test all distilled models with the same 4-step sampling, which is shown to be able to achieve the best sampling performance; we then compare each distilled model with the undistilled one on the normalized sampling cost. We follow the same practice discussed before for selecting the optimal CFG rate and compare them under the same relative inference cost. The results shown in the left part of Fig. 2.13 demonstrate that distillation significantly improves the generative performance for all

models in 4-step sampling, with FID improvements across the board. By comparing these distilled models with the undistilled models in the right part of Fig. 2.13, we demonstrate that distilled models outperform undistilled models at the same sampling cost. However, at the specific sampling cost, *i.e.*, sampling cost $\approx$ 8, the smaller undistilled 83M model still achieves similar performance to the larger distilled 866M model. The observation further supports our proposed scaling sampling-efficiency after diffusion distillation.

2.3 Conclusion

In this paper, we investigated scaling properties of Latent Diffusion Models (LDMs), specifically through scaling model size from 39 million to 5 billion parameters. We trained these scaled models from scratch on a web-scale text-to-image dataset and then finetuned the pretrained models for downstream tasks. Our findings unveil that, under identical sampling costs, smaller models frequently outperform larger models, suggesting a promising direction for accelerating LDMs in terms of model size. We further show that the sampling efficiency is consistent in multiple axes. For example, it is invariant to various diffusion samplers (stochastic and deterministic), and also holds true for distilled models. We believe this analysis of scaling sampling efficiency would be instrumental in guiding future developments of LDMs, specifically for balancing model size against performance and efficiency in a broad spectrum of practical applications.

Chapter 3
Chapter title goes here

Since it is written in LaTeX, it must be true.

– Isaac Newton

3.1 Introduction

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special contents, but the length of words should match the language.

3.2 Current approach

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written

in of the original language. There is no need for special contents, but the length of words should match the language. [1]

3.2.1 Hypothesis statement

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special contents, but the length of words should match the language. [2]

3.2.2 Experimental evidences

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special contents, but the length of

[1] Hello, this is the first footnote with no indentation and single-spaced text. The spacing between two footnotes is also single-spaced.

[2] This is the second footnote.

words should match the language.

3.2.2.1 Data analysis

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special contents, but the length of words should match the language.

Chapter 4

A long chapter title test the distance to the quote

This is a large quote placed with single spacing over two lines

– Unknown Author

4.1 A long section heading to test the distance before this

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special contents, but the length of words should match the language.

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special contents, but the length of

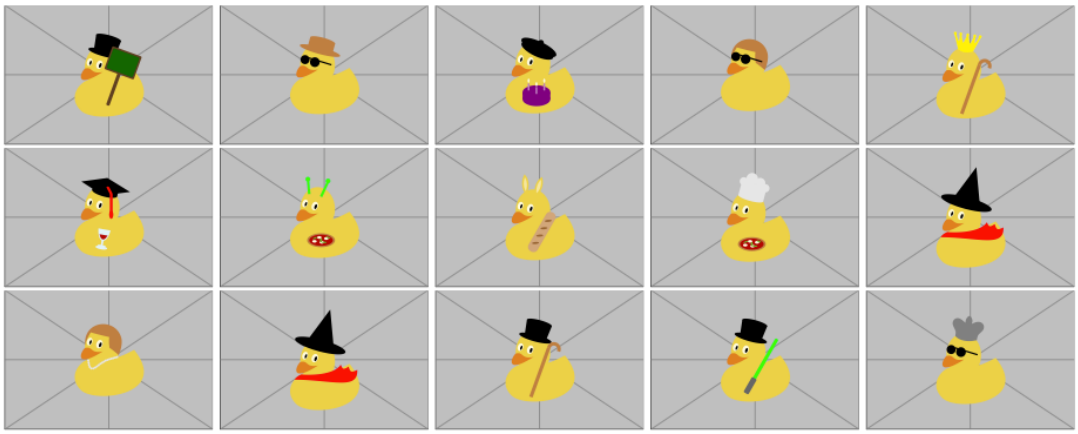


Figure 4.1: Here are some photos of ducks to make you feel happy in tough times.

words should match the language.

This is the second paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special contents, but the length of words should match the language.

Chapter 5

Chapter title goes here

5.1 Introduction

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special contents, but the length of words should match the language.

Col1	Col2	Col2	Col3
1	6	87837	787
2	7	78	5415
3	545	778	7507
4	545	18744	7560
5	88	788	6344

Table 5.1: Table to test captions and labels taken from Overleaf.

And after the second paragraph follows the third paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at

this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special contents, but the length of words should match the language.

After this fourth paragraph, we start a new paragraph sequence. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special contents, but the length of words should match the language.

Bibliographic references

1. Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S. & Poole, B. *Score-based generative modeling through stochastic differential equations* in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021* (2021).

2. Ho, J., Jain, A. & Abbeel, P. *Denoising diffusion probabilistic models* in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual* (eds Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. & Lin, H.) (2020).

3. Kingma, D. P. & Welling, M. *Auto-encoding variational bayes* in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings* (eds Bengio, Y. & LeCun, Y.) (2014).

4. Song, Y., Garg, S., Shi, J. & Ermon, S. *Sliced score matching: A scalable approach to density and score estimation* in *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2019, Tel Aviv, Israel, July 22-25, 2019* (eds Globerson, A. & Silva, R.) (2019).

5. Kingma, D., Salimans, T., Poole, B. & Ho, J. Variational diffusion models. *Advances in neural information processing systems* (2021).

6. Song, J., Meng, C. & Ermon, S. *Denoising diffusion implicit models* in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021* (2021).

7. Dhariwal, P. & Nichol, A. Q. *Diffusion models beat gans on image synthesis* in *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual* (eds Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P. & Vaughan, J. W.) (2021).

8. Salimans, T. & Ho, J. *Progressive distillation for fast sampling of diffusion models* in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022* (2022).

9. 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* (2022).

10. 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 *ECCV* (2014).

11. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., *et al.* Laion-5b: an open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems* (2022).

12. 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* (2023).

13. Delbracio, M. & Milanfar, P. Inversion by direct iteration: an alternative to denoising diffusion for image restoration. *Transactions on Machine Learning Research.* Featured Certification (2023).

14. Ren, M., Delbracio, M., Talebi, H., Gerig, G. & Milanfar, P. *Multiscale structure guided diffusion for image deblurring* in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), 10721–10733.

15. Qi, C., Tu, Z., Ye, K., Delbracio, M., Milanfar, P., Chen, Q. & Talebi, H. Tip: text-driven image processing with semantic and restoration instructions. *ArXiv preprint* (2023).

16. Mei, K. & Patel, V. *Vidm: video implicit diffusion models* in *Proceedings of the AAAI Conference on Artificial Intelligence* **37** (2023), 9117–9125.

17. Mei, K., Zhou, M. & Patel, V. M. T1: scaling diffusion probabilistic fields to high-resolution on unified visual modalities. *ArXiv preprint* (2023).

18. Wu, J. Z., Ge, Y., Wang, X., Lei, S. W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X. & Shou, M. Z. *Tune-a-video: one-shot tuning of image diffusion models for text-to-video generation* in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023).

19. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., *et al.* Make-a-video: text-to-video generation without text-video data. *ArXiv preprint* (2022).

20. Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D., Wang, W. & Plumbley, M. D. Audioldm: text-to-audio generation with latent diffusion models. *ArXiv preprint* (2023).

21. Lin, C.-H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.-Y. & Lin, T.-Y. *Magic3d: high-resolution text-to-3d content creation* in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023).

22. Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S. & Vondrick, C. *Zero-1-to-3: zero-shot one image to 3d object* in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023).

23. Du, H., Zhang, R., Niyato, D., Kang, J., Xiong, Z., Kim, D. I., Shen, X. S. & Poor, H. V. Exploring collaborative distributed diffusion-based ai-generated content (aigc) in wireless networks. *IEEE Network* (2023).

24. 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* (2023).

25. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. & Bengio, Y. Generative adversarial networks. *Communications of the ACM* (2020).

26. Karras, T., Laine, S. & Aila, T. *A style-based generator architecture for generative adversarial networks* in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019* (2019).

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. *NeurIPS* (2023).

28. Zhao, Y., Xu, Y., Xiao, Z. & Hou, T. Mobilediffusion: subsecond text-to-image generation on mobile devices. *ArXiv preprint* (2023).

29. Peebles, W. & Xie, S. *Scalable diffusion models with transformers* in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023).

30. 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).

31. Kim, B.-K., Song, H.-K., Castells, T. & Choi, S. On architectural compression of text-to-image diffusion models. *ArXiv preprint* (2023).

32. Dockhorn, T., Vahdat, A. & Kreis, K. Genie: higher-order denoising diffusion solvers. *Advances in Neural Information Processing Systems* (2022).

33. Karras, T., Aittala, M., Aila, T. & Laine, S. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems* (2022).

34. 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* (2022).

35. 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* (2023).

36. Xu, Y., Zhao, Y., Xiao, Z. & Hou, T. Ufogen: you forward once large scale text-to-image generation via diffusion gans. *ArXiv preprint* (2023).

37. Luhman, E. & Luhman, T. Knowledge distillation in iterative generative models for improved sampling speed. *ArXiv preprint* (2021).

38. Song, Y., Dhariwal, P., Chen, M. & Sutskever, I. Consistency models. *ICML* (2023).

39. Sauer, A., Lorenz, D., Blattmann, A. & Rombach, R. Adversarial diffusion distillation. *ArXiv preprint* (2023).

40. Gu, J., Zhai, S., Zhang, Y., Liu, L. & Susskind, J. M. *Boot: data-free distillation of denoising diffusion models with bootstrapping* in *ICML 2023 Workshop on Structured Probabilistic Inference {\&} Generative Modeling* (2023).

41. Mei, K., Delbracio, M., Talebi, H., Tu, Z., Patel, V. M. & Milanfar, P. *Codi: conditional diffusion distillation for higher-fidelity and faster image generation* in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024).

42. Jouppi, N., Kurian, G., Li, S., Ma, P., Nagarajan, R., Nai, L., Patil, N., Subramanian, S., Swing, A., Towles, B., *et al. Tpu v4: an optically reconfigurable supercomputer for machine learning with hardware support for embeddings* in *Proceedings of the 50th Annual International Symposium on Computer Architecture* (2023).

43. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D. J. & Norouzi, M. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

44. Sahak, H., Watson, D., Saharia, C. & Fleet, D. Denoising diffusion probabilistic models for robust image super-resolution in the wild. *ArXiv preprint* (2023).

45. 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* (2023).

46. 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* (2022).

47. Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyer, L., *et al.* Pali: a jointly-scaled multilingual language-image model. *ArXiv preprint* (2022).

48. Wang, X., Xie, L., Dong, C. & Shan, Y. *Real-esrgan: training real-world blind super-resolution with pure synthetic data* in *IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2021, Montreal, BC, Canada, October 11-17, 2021* (2021).

49. Ho, J. & Salimans, T. Classifier-free diffusion guidance. *ArXiv preprint* (2022).

50. Luo, S., Tan, Y., Huang, L., Li, J. & Zhao, H. Latent consistency models: synthesizing high-resolution images with few-step inference. *ArXiv preprint* (2023).

51. Lin, S., Wang, A. & Yang, X. Sdxl-lightning: progressive adversarial diffusion distillation. *ArXiv preprint* (2024).

Appendix A
Some necessary information

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special contents, but the length of words should match the language.

This is the second paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special contents, but the length of words should match the language.

And after the second paragraph follows the third paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no

information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special contents, but the length of words should match the language.

Appendix B

A few more additional information

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special contents, but the length of words should match the language.