High-Resolution Image Synthesis with Latent Diffusion Models

Robin Rombach[1] * Andreas Blattmann[1] * Dominik Lorenz[1] Patrick Esser[R] Björn Ommer[1]
[1]Ludwig Maximilian University of Munich & IWR, Heidelberg University, Germany [R]Runway ML
https://github.com/CompVis/latent-diffusion

Abstract

By decomposing the image formation process into a sequential application of denoising autoencoders, diffusion models (DMs) achieve state-of-the-art synthesis results on image data and beyond. Additionally, their formulation allows for a guiding mechanism to control the image generation process without retraining. However, since these models typically operate directly in pixel space, optimization of powerful DMs often consumes hundreds of GPU days and inference is expensive due to sequential evaluations. To enable DM training on limited computational resources while retaining their quality and flexibility, we apply them in the latent space of powerful pretrained autoencoders. In contrast to previous work, training diffusion models on such a representation allows for the first time to reach a near-optimal point between complexity reduction and detail preservation, greatly boosting visual fidelity. By introducing cross-attention layers into the model architecture, we turn diffusion models into powerful and flexible generators for general conditioning inputs such as text or bounding boxes and high-resolution synthesis becomes possible in a convolutional manner. Our latent diffusion models (LDMs) achieve new state-of-the-art scores for image inpainting and class-conditional image synthesis and highly competitive performance on various tasks, including text-to-image synthesis, unconditional image generation and super-resolution, while significantly reducing computational requirements compared to pixel-based DMs.

1. Introduction

Image synthesis is one of the computer vision fields with the most spectacular recent development, but also among those with the greatest computational demands. Especially high-resolution synthesis of complex, natural scenes is presently dominated by scaling up likelihood-based models, potentially containing billions of parameters in autoregressive (AR) transformers [66,67]. In contrast, the promising results of GANs [3, 27, 40] have been revealed to be mostly confined to data with comparably limited variability as their adversarial learning procedure does not easily scale to modeling complex, multi-modal distributions. Recently, diffusion models [82], which are built from a hierarchy of denoising autoencoders, have shown to achieve impressive results in image synthesis [30,85] and beyond [7,45,48,57], and define the state-of-the-art in class-conditional image synthesis [15,31] and super-resolution [72]. Moreover, even unconditional DMs can readily be applied to tasks such as inpainting and colorization [85] or stroke-based synthesis [53], in contrast to other types of generative models [19,46,69]. Being likelihood-based models, they do not exhibit mode-collapse and training instabilities as GANs and, by heavily exploiting parameter sharing, they can model highly complex distributions of natural images without involving billions of parameters as in AR models [67].

Figure 1. Boosting the upper bound on achievable quality with less agressive downsampling. Since diffusion models offer excellent inductive biases for spatial data, we do not need the heavy spatial downsampling of related generative models in latent space, but can still greatly reduce the dimensionality of the data via suitable autoencoding models, see Sec. 3. Images are from the DIV2K [1] validation set, evaluated at 512^2 px. We denote the spatial downsampling factor by f. Reconstruction FIDs [29] and PSNR are calculated on ImageNet-val. [12]; see also Tab. 8.

Democratizing High-Resolution Image Synthesis DMs belong to the class of likelihood-based models, whose mode-covering behavior makes them prone to spend excessive amounts of capacity (and thus compute resources) on modeling imperceptible details of the data [16, 73]. Although the reweighted variational objective [30] aims to address this by undersampling the initial denoising steps, DMs are still computationally demanding, since training and evaluating such a model requires repeated function evaluations (and gradient computations) in the high-dimensional space of RGB images. As an example, training the most powerful DMs often takes hundreds of GPU days (*e.g.* 150 - 1000 V100 days in [15]) and repeated evaluations on a noisy version of the input space render also inference expensive,

*The first two authors contributed equally to this work.

so that producing 50k samples takes approximately 5 days [15] on a single A100 GPU. This has two consequences for the research community and users in general: Firstly, training such a model requires massive computational resources only available to a small fraction of the field, and leaves a huge carbon footprint [65, 86]. Secondly, evaluating an already trained model is also expensive in time and memory, since the same model architecture must run sequentially for a large number of steps (*e.g.* 25 - 1000 steps in [15]).

To increase the accessibility of this powerful model class and at the same time reduce its significant resource consumption, a method is needed that reduces the computational complexity for both training and sampling. Reducing the computational demands of DMs without impairing their performance is, therefore, key to enhance their accessibility.

Departure to Latent Space Our approach starts with the analysis of already trained diffusion models in pixel space: Fig. 2 shows the rate-distortion trade-off of a trained model. As with any likelihood-based model, learning can be roughly divided into two stages: First is a *perceptual compression* stage which removes high-frequency details but still learns little semantic variation. In the second stage, the actual generative model learns the semantic and conceptual composition of the data (*semantic compression*). We thus aim to first find a *perceptually equivalent, but computationally more suitable space*, in which we will train diffusion models for high-resolution image synthesis.

Following common practice [11, 23, 66, 67, 96], we separate training into two distinct phases: First, we train an autoencoder which provides a lower-dimensional (and thereby efficient) representational space which is perceptually equivalent to the data space. Importantly, and in contrast to previous work [23, 66], we do not need to rely on excessive spatial compression, as we train DMs in the learned latent space, which exhibits better scaling properties with respect to the spatial dimensionality. The reduced complexity also provides efficient image generation from the latent space with a single network pass. We dub the resulting model class *Latent Diffusion Models* (LDMs).

A notable advantage of this approach is that we need to train the universal autoencoding stage only once and can therefore reuse it for multiple DM trainings or to explore possibly completely different tasks [81]. This enables efficient exploration of a large number of diffusion models for various image-to-image and text-to-image tasks. For the latter, we design an architecture that connects transformers to the DM's UNet backbone [71] and enables arbitrary types of token-based conditioning mechanisms, see Sec. 3.3.

In sum, our work makes the following **contributions**:

(i) In contrast to purely transformer-based approaches [23, 66], our method scales more graceful to higher dimensional data and can thus (a) work on a compression level which provides more faithful and detailed reconstructions than previous work (see Fig. 1) and (b) can be efficiently applied to high-resolution synthesis of megapixel images.

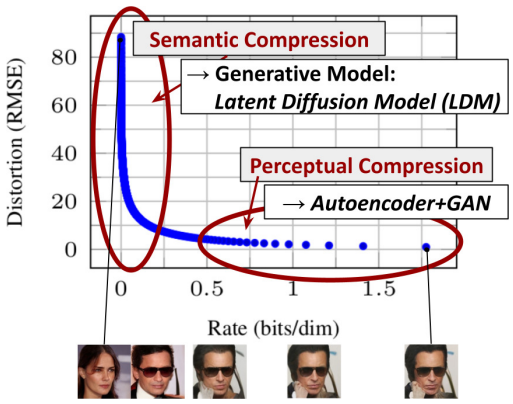

Figure 2. Illustrating perceptual and semantic compression: Most bits of a digital image correspond to imperceptible details. While DMs allow to suppress this semantically meaningless information by minimizing the responsible loss term, gradients (during training) and the neural network backbone (training and inference) still need to be evaluated on all pixels, leading to superfluous computations and unnecessarily expensive optimization and inference. We propose *latent diffusion models (LDMs)* as an effective generative model and a separate mild compression stage that only eliminates imperceptible details. Data and images from [30].

(ii) We achieve competitive performance on multiple tasks (unconditional image synthesis, inpainting, stochastic super-resolution) and datasets while significantly lowering computational costs. Compared to pixel-based diffusion approaches, we also significantly decrease inference costs.

(iii) We show that, in contrast to previous work [93] which learns both an encoder/decoder architecture and a score-based prior simultaneously, our approach does not require a delicate weighting of reconstruction and generative abilities. This ensures extremely faithful reconstructions and requires very little regularization of the latent space.

(iv) We find that for densely conditioned tasks such as super-resolution, inpainting and semantic synthesis, our model can be applied in a convolutional fashion and render large, consistent images of $\sim 1024^2$ px.

(v) Moreover, we design a general-purpose conditioning mechanism based on cross-attention, enabling multi-modal training. We use it to train class-conditional, text-to-image and layout-to-image models.

(vi) Finally, we release pretrained latent diffusion and autoencoding models at `https://github.com/CompVis/latent-diffusion` which might be reusable for a various tasks besides training of DMs [81].

2. Related Work

Generative Models for Image Synthesis The high dimensional nature of images presents distinct challenges to generative modeling. Generative Adversarial Networks (GAN) [27] allow for efficient sampling of high resolution images with good perceptual quality [3, 42], but are diffi-

cult to optimize [2, 28, 54] and struggle to capture the full data distribution [55]. In contrast, likelihood-based methods emphasize good density estimation which renders optimization more well-behaved. Variational autoencoders (VAE) [46] and flow-based models [18, 19] enable efficient synthesis of high resolution images [9, 44, 92], but sample quality is not on par with GANs. While autoregressive models (ARM) [6, 10, 94, 95] achieve strong performance in density estimation, computationally demanding architectures [97] and a sequential sampling process limit them to low resolution images. Because pixel based representations of images contain barely perceptible, high-frequency details [16,73], maximum-likelihood training spends a disproportionate amount of capacity on modeling them, resulting in long training times. To scale to higher resolutions, several two-stage approaches [23,67,101,103] use ARMs to model a compressed latent image space instead of raw pixels.

Recently, **Diffusion Probabilistic Models** (DM) [82], have achieved state-of-the-art results in density estimation [45] as well as in sample quality [15]. The generative power of these models stems from a natural fit to the inductive biases of image-like data when their underlying neural backbone is implemented as a UNet [15, 30, 71, 85]. The best synthesis quality is usually achieved when a reweighted objective [30] is used for training. In this case, the DM corresponds to a lossy compressor and allow to trade image quality for compression capabilities. Evaluating and optimizing these models in pixel space, however, has the downside of low inference speed and very high training costs. While the former can be partially adressed by advanced sampling strategies [47, 75, 84] and hierarchical approaches [31, 93], training on high-resolution image data always requires to calculate expensive gradients. We adress both drawbacks with our proposed *LDMs*, which work on a compressed latent space of lower dimensionality. This renders training computationally cheaper and speeds up inference with almost no reduction in synthesis quality (see Fig. 1).

Two-Stage Image Synthesis To mitigate the shortcomings of individual generative approaches, a lot of research [11, 23, 67, 70, 101, 103] has gone into combining the strengths of different methods into more efficient and performant models via a two stage approach. VQ-VAEs [67, 101] use autoregressive models to learn an expressive prior over a discretized latent space. [66] extend this approach to text-to-image generation by learning a joint distributation over discretized image and text representations. More generally, [70] uses conditionally invertible networks to provide a generic transfer between latent spaces of diverse domains. Different from VQ-VAEs, VQGANs [23, 103] employ a first stage with an adversarial and perceptual objective to scale autoregressive transformers to larger images. However, the high compression rates required for feasible ARM training, which introduces billions of trainable parameters [23, 66], limit the overall performance of such approaches and less compression comes at the price of high computational cost [23, 66]. Our work prevents such trade-offs, as our proposed *LDMs* scale more gently to higher dimensional latent spaces due to their convolutional backbone. Thus, we are free to choose the level of compression which optimally mediates between learning a powerful first stage, without leaving too much perceptual compression up to the generative diffusion model while guaranteeing high-fidelity reconstructions (see Fig. 1).

While approaches to jointly [93] or separately [80] learn an encoding/decoding model together with a score-based prior exist, the former still require a difficult weighting between reconstruction and generative capabilities [11] and are outperformed by our approach (Sec. 4), and the latter focus on highly structured images such as human faces.

3. Method

To lower the computational demands of training diffusion models towards high-resolution image synthesis, we observe that although diffusion models allow to ignore perceptually irrelevant details by undersampling the corresponding loss terms [30], they still require costly function evaluations in pixel space, which causes huge demands in computation time and energy resources.

We propose to circumvent this drawback by introducing an explicit separation of the compressive from the generative learning phase (see Fig. 2). To achieve this, we utilize an autoencoding model which learns a space that is perceptually equivalent to the image space, but offers significantly reduced computational complexity.

Such an approach offers several advantages: (i) By leaving the high-dimensional image space, we obtain DMs which are computationally much more efficient because sampling is performed on a low-dimensional space. (ii) We exploit the inductive bias of DMs inherited from their UNet architecture [71], which makes them particularly effective for data with spatial structure and therefore alleviates the need for aggressive, quality-reducing compression levels as required by previous approaches [23, 66]. (iii) Finally, we obtain general-purpose compression models whose latent space can be used to train multiple generative models and which can also be utilized for other downstream applications such as single-image CLIP-guided synthesis [25].

3.1. Perceptual Image Compression

Our perceptual compression model is based on previous work [23] and consists of an autoencoder trained by combination of a perceptual loss [106] and a patch-based [33] adversarial objective [20, 23, 103]. This ensures that the reconstructions are confined to the image manifold by enforcing local realism and avoids bluriness introduced by relying solely on pixel-space losses such as L_2 or L_1 objectives.

More precisely, given an image $x \in \mathbb{R}^{H\times W\times 3}$ in RGB space, the encoder $\mathcal{E}$ encodes x into a latent representa-

tion $z = \mathcal{E}(x)$, and the decoder $\mathcal{D}$ reconstructs the image from the latent, giving $\tilde{x} = \mathcal{D}(z) = \mathcal{D}(\mathcal{E}(x))$, where $z \in \mathbb{R}^{h \times w \times c}$. Importantly, the encoder *downsamples* the image by a factor $f = H/h = W/w$, and we investigate different downsampling factors $f = 2^m$, with $m \in \mathbb{N}$.

In order to avoid arbitrarily high-variance latent spaces, we experiment with two different kinds of regularizations. The first variant, *KL-reg.*, imposes a slight KL-penalty towards a standard normal on the learned latent, similar to a VAE [46, 69], whereas *VQ-reg.* uses a vector quantization layer [96] within the decoder. This model can be interpreted as a VQGAN [23] but with the quantization layer absorbed by the decoder. Because our subsequent DM is designed to work with the two-dimensional structure of our learned latent space $z = \mathcal{E}(x)$, we can use relatively mild compression rates and achieve very good reconstructions. This is in contrast to previous works [23, 66], which relied on an arbitrary 1D ordering of the learned space z to model its distribution autoregressively and thereby ignored much of the inherent structure of z. Hence, our compression model preserves details of x better (see Tab. 8). The full objective and training details can be found in the supplement.

3.2. Latent Diffusion Models

Diffusion Models [82] are probabilistic models designed to learn a data distribution $p(x)$ by gradually denoising a normally distributed variable, which corresponds to learning the reverse process of a fixed Markov Chain of length T. For image synthesis, the most successful models [15, 30, 72] rely on a reweighted variant of the variational lower bound on $p(x)$, which mirrors denoising score-matching [85]. These models can be interpreted as an equally weighted sequence of denoising autoencoders $\epsilon_\theta(x_t, t)$; $t = 1 \dots T$, which are trained to predict a denoised variant of their input x_t, where x_t is a noisy version of the input x. The corresponding objective can be simplified to (Sec. B)

$$L_{DM} = \mathbb{E}_{x, \epsilon \sim \mathcal{N}(0,1), t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\right], \qquad (1)$$

with t uniformly sampled from $\{1, \dots, T\}$.

Generative Modeling of Latent Representations With our trained perceptual compression models consisting of $\mathcal{E}$ and $\mathcal{D}$, we now have access to an efficient, low-dimensional latent space in which high-frequency, imperceptible details are abstracted away. Compared to the high-dimensional pixel space, this space is more suitable for likelihood-based generative models, as they can now (i) focus on the important, semantic bits of the data and (ii) train in a lower dimensional, computationally much more efficient space.

Unlike previous work that relied on autoregressive, attention-based transformer models in a highly compressed, discrete latent space [23, 66, 103], we can take advantage of image-specific inductive biases that our model offers. This includes the ability to build the underlying UNet primarily from 2D convolutional layers, and further focusing the objective on the perceptually most relevant bits using the reweighted bound, which now reads

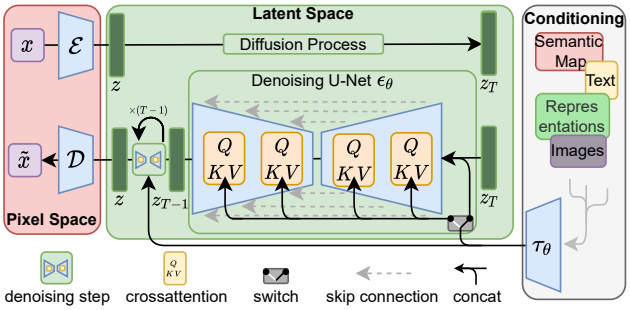

Figure 3. We condition LDMs either via concatenation or by a more general cross-attention mechanism. See Sec. 3.3

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x), \epsilon \sim \mathcal{N}(0,1), t}\left[\|\epsilon - \epsilon_\theta(z_t, t)\|_2^2\right]. \qquad (2)$$

The neural backbone $\epsilon_\theta(\circ, t)$ of our model is realized as a time-conditional UNet [71]. Since the forward process is fixed, z_t can be efficiently obtained from $\mathcal{E}$ during training, and samples from $p(z)$ can be decoded to image space with a single pass through $\mathcal{D}$.

3.3. Conditioning Mechanisms

Similar to other types of generative models [56, 83], diffusion models are in principle capable of modeling conditional distributions of the form $p(z|y)$. This can be implemented with a conditional denoising autoencoder $\epsilon_\theta(z_t, t, y)$ and paves the way to controlling the synthesis process through inputs y such as text [68], semantic maps [33, 61] or other image-to-image translation tasks [34].

In the context of image synthesis, however, combining the generative power of DMs with other types of conditionings beyond class-labels [15] or blurred variants of the input image [72] is so far an under-explored area of research.

We turn DMs into more flexible conditional image generators by augmenting their underlying UNet backbone with the cross-attention mechanism [97], which is effective for learning attention-based models of various input modalities [35, 36]. To pre-process y from various modalities (such as language prompts) we introduce a domain specific encoder τ_θ that projects y to an intermediate representation $\tau_\theta(y) \in \mathbb{R}^{M \times d_\tau}$, which is then mapped to the intermediate layers of the UNet via a cross-attention layer implementing $\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V$, with

$$Q = W_Q^{(i)} \cdot \varphi_i(z_t), \; K = W_K^{(i)} \cdot \tau_\theta(y), \; V = W_V^{(i)} \cdot \tau_\theta(y).$$

Here, $\varphi_i(z_t) \in \mathbb{R}^{N \times d_\epsilon^i}$ denotes a (flattened) intermediate representation of the UNet implementing ϵ_θ and $W_V^{(i)} \in$

Figure 4. Samples from *LDMs* trained on CelebAHQ [39], FFHQ [41], LSUN-Churches [102], LSUN-Bedrooms [102] and class-conditional ImageNet [12], each with a resolution of 256×256. Best viewed when zoomed in. For more samples *cf*. the supplement.

$\mathbb{R}^{d \times d_\epsilon^i}$, $W_Q^{(i)} \in \mathbb{R}^{d \times d_\tau}$ & $W_K^{(i)} \in \mathbb{R}^{d \times d_\tau}$ are learnable projection matrices [36, 97]. See Fig. 3 for a visual depiction.

Based on image-conditioning pairs, we then learn the conditional LDM via

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|_2^2\right], \quad (3)$$

where both τ_θ and ϵ_θ are jointly optimized via Eq. 3. This conditioning mechanism is flexible as τ_θ can be parameterized with domain-specific experts, *e.g.* (unmasked) transformers [97] when y are text prompts (see Sec. 4.3.1)

4. Experiments

LDMs provide means to flexible and computationally tractable diffusion based image synthesis of various image modalities, which we empirically show in the following. Firstly, however, we analyze the gains of our models compared to pixel-based diffusion models in both training and inference. Interestingly, we find that *LDMs* trained in *VQ*-regularized latent spaces sometimes achieve better sample quality, even though the reconstruction capabilities of *VQ*-regularized first stage models slightly fall behind those of their continuous counterparts, *cf*. Tab. 8. A visual comparison between the effects of first stage regularization schemes on *LDM* training and their generalization abilities to resolutions $> 256^2$ can be found in Appendix D.1. In E.2 we list details on architecture, implementation, training and evaluation for all results presented in this section.

4.1. On Perceptual Compression Tradeoffs

This section analyzes the behavior of our LDMs with different downsampling factors $f \in \{1, 2, 4, 8, 16, 32\}$ (abbreviated as *LDM-f*, where *LDM-1* corresponds to pixel-based DMs). To obtain a comparable test-field, we fix the computational resources to a single NVIDIA A100 for all experiments in this section and train all models for the same number of steps and with the same number of parameters.

Tab. 8 shows hyperparameters and reconstruction performance of the first stage models used for the *LDMs* compared in this section. Fig. 6 shows sample quality as a function of training progress for 2M steps of class-conditional models on the ImageNet [12] dataset. We see that, i) small downsampling factors for *LDM-{1,2}* result in slow training progress, whereas ii) overly large values of f cause stagnating fidelity after comparably few training steps. Revisiting the analysis above (Fig. 1 and 2) we attribute this to i) leaving most of perceptual compression to the diffusion model and ii) too strong first stage compression resulting in information loss and thus limiting the achievable quality. *LDM-{4-16}* strike a good balance between efficiency and perceptually faithful results, which manifests in a significant FID [29] gap of 38 between pixel-based diffusion (*LDM-1*) and *LDM-8* after 2M training steps.

In Fig. 7, we compare models trained on CelebA-HQ [39] and ImageNet in terms sampling speed for different numbers of denoising steps with the DDIM sampler [84] and plot it against FID-scores [29]. *LDM-{4-8}* outperform models with unsuitable ratios of perceptual and conceptual compression. Especially compared to pixel-based *LDM-1*, they achieve much lower FID scores while simultaneously significantly increasing sample throughput. Complex datasets such as ImageNet require reduced compression rates to avoid reducing quality. In summary, *LDM-4* and *-8* offer the best conditions for achieving high-quality synthesis results.

4.2. Image Generation with Latent Diffusion

We train unconditional models of 256^2 images on CelebA-HQ [39], FFHQ [41], LSUN-Churches and -Bedrooms [102] and evaluate the i) sample quality and ii) their coverage of the data manifold using ii) FID [29] and ii) Precision-and-Recall [50]. Tab. 1 summarizes our results. On CelebA-HQ, we report a new state-of-the-art FID of 5.11, outperforming previous likelihood-based models as well as GANs. We also outperform LSGM [93] where a latent diffusion model is trained jointly together with the first stage. In contrast, we train diffusion models in a fixed space

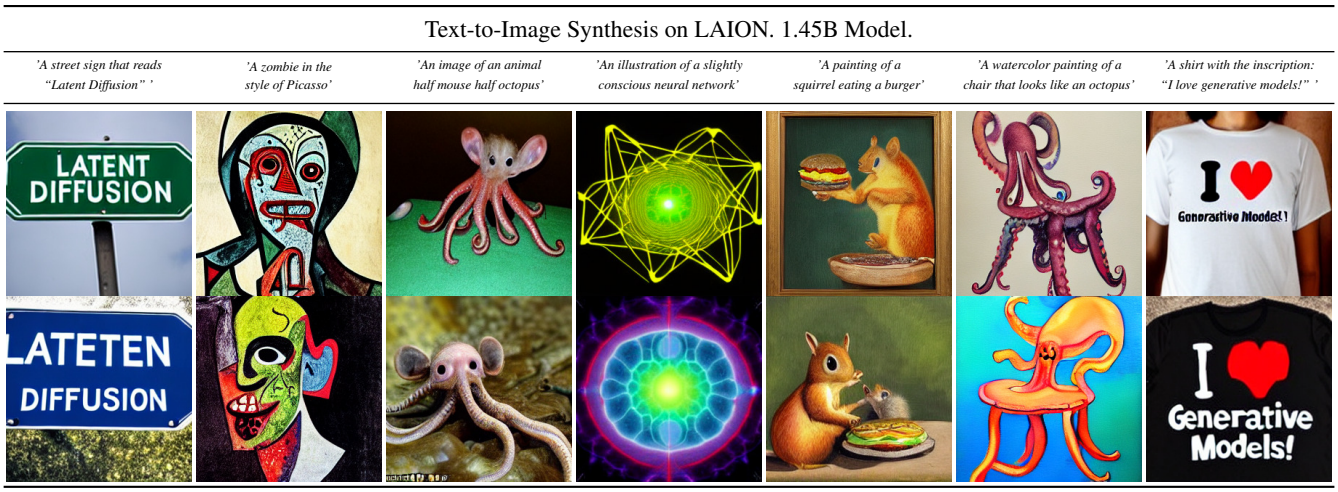

Figure 5. Samples for user-defined text prompts from our model for text-to-image synthesis, *LDM-8 (KL)*, which was trained on the LAION [78] database. Samples generated with 200 DDIM steps and $\eta = 1.0$. We use unconditional guidance [32] with $s = 10.0$.

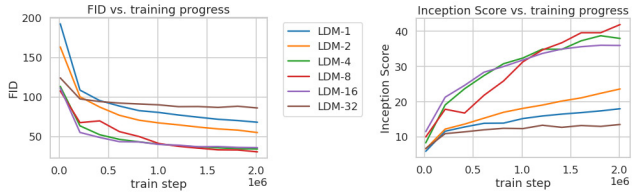

Figure 6. Analyzing the training of class-conditional *LDMs* with different downsampling factors f over 2M train steps on the ImageNet dataset. Pixel-based *LDM-1* requires substantially larger train times compared to models with larger downsampling factors (*LDM-*$\{$*4-16*$\}$). Too much perceptual compression as in *LDM-32* limits the overall sample quality. All models are trained on a single NVIDIA A100 with the same computational budget. Results obtained with 100 DDIM steps [84] and $\kappa = 0$.

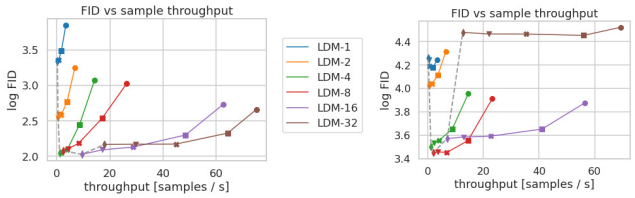

Figure 7. Comparing *LDMs* with varying compression on the CelebA-HQ (left) and ImageNet (right) datasets. Different markers indicate $\{10, 20, 50, 100, 200\}$ sampling steps using DDIM, from right to left along each line. The dashed line shows the FID scores for 200 steps, indicating the strong performance of *LDM-*$\{$*4-8*$\}$. FID scores assessed on 5000 samples. All models were trained for 500k (CelebA) / 2M (ImageNet) steps on an A100.

and avoid the difficulty of weighing reconstruction quality against learning the prior over the latent space, see Fig. 1-2.

We outperform prior diffusion based approaches on all but the LSUN-Bedrooms dataset, where our score is close to ADM [15], despite utilizing half its parameters and requiring 4-times less train resources (see Appendix E.3.5).

CelebA-HQ 256 × 256				FFHQ 256 × 256			
Method	FID ↓	Prec. ↑	Recall ↑	**Method**	FID ↓	Prec. ↑	Recall ↑
DC-VAE [63]	15.8	-	-	ImageBART [21]	9.57	-	-
VQGAN+T. [23] (k=400)	10.2	-	-	U-Net GAN (+aug) [77]	10.9 (7.6)	-	-
PGGAN [39]	8.0	-	-	UDM [43]	5.54	-	-
LSGM [93]	7.22	-	-	StyleGAN [41]	4.16	0.71	0.46
UDM [43]	7.16	-	-	ProjectedGAN [76]	**3.08**	0.65	0.46
LDM-4 (ours, 500-s†)	**5.11**	0.72	0.49	*LDM-4* (ours, 200-s)	4.98	**0.73**	**0.50**

LSUN-Churches 256 × 256				LSUN-Bedrooms 256 × 256			
Method	FID ↓	Prec. ↑	Recall ↑	**Method**	FID ↓	Prec. ↑	Recall ↑
DDPM [30]	7.89	-	-	ImageBART [21]	5.51	-	-
ImageBART [21]	7.32	-	-	DDPM [30]	4.9	-	-
PGGAN [39]	6.42	-	-	UDM [43]	4.57	-	-
StyleGAN [41]	4.21	-	-	StyleGAN [41]	2.35	0.59	0.48
StyleGAN2 [42]	3.86	-	-	ADM [15]	1.90	**0.66**	**0.51**
ProjectedGAN [76]	**1.59**	0.61	0.44	ProjectedGAN [76]	**1.52**	0.61	0.34
*LDM-8** (ours, 200-s)	4.02	**0.64**	**0.52**	*LDM-4* (ours, 200-s)	2.95	**0.66**	0.48

Table 1. Evaluation metrics for unconditional image synthesis. CelebA-HQ results reproduced from [43, 63, 100], FFHQ from [42, 43]. †: N-s refers to N sampling steps with the DDIM [84] sampler. *: trained in *KL*-regularized latent space. Additional results can be found in the supplementary.

Text-Conditional Image Synthesis				
Method	FID ↓	IS↑	Nparams	
CogView† [17]	27.10	18.20	4B	self-ranking, rejection rate 0.017
LAFITE† [109]	26.94	26.02	75M	
GLIDE* [59]	12.24	-	6B	277 DDIM steps, c.f.g. [32] $s = 3$
Make-A-Scene* [26]	**11.84**	-	4B	c.f.g for AR models [98] $s = 5$
LDM-KL-8	23.31	$20.03_{\pm 0.33}$	1.45B	250 DDIM steps
*LDM-KL-8-G**	12.63	$\mathbf{30.29}_{\pm \mathbf{0.42}}$	1.45B	250 DDIM steps, c.f.g. [32] $s = 1.5$

Table 2. Evaluation of text-conditional image synthesis on the 256 × 256-sized MS-COCO [51] dataset: with 250 DDIM [84] steps our model is on par with the most recent diffusion [59] and autoregressive [26] methods despite using significantly less parameters. †/*:Numbers from [109]/ [26]

Moreover, *LDMs* consistently improve upon GAN-based methods in Precision and Recall, thus confirming the advantages of their mode-covering likelihood-based training objective over adversarial approaches. In Fig. 4 we also show qualitative results on each dataset.

Figure 8. Layout-to-image synthesis with an *LDM* on COCO [4], see Sec. 4.3.1. Quantitative evaluation in the supplement D.3.

4.3. Conditional Latent Diffusion

4.3.1 Transformer Encoders for LDMs

By introducing cross-attention based conditioning into LDMs we open them up for various conditioning modalities previously unexplored for diffusion models. For **text-to-image** image modeling, we train a 1.45B parameter *KL*-regularized *LDM* conditioned on language prompts on LAION-400M [78]. We employ the BERT-tokenizer [14] and implement τ_θ as a transformer [97] to infer a latent code which is mapped into the UNet via (multi-head) cross-attention (Sec. 3.3). This combination of domain specific experts for learning a language representation and visual synthesis results in a powerful model, which generalizes well to complex, user-defined text prompts, *cf*. Fig. 8 and 5. For quantitative analysis, we follow prior work and evaluate text-to-image generation on the MS-COCO [51] validation set, where our model improves upon powerful AR [17, 66] and GAN-based [109] methods, *cf*. Tab. 2. We note that applying classifier-free diffusion guidance [32] greatly boosts sample quality, such that the guided *LDM-KL-8-G* is on par with the recent state-of-the-art AR [26] and diffusion models [59] for text-to-image synthesis, while substantially reducing parameter count. To further analyze the flexibility of the cross-attention based conditioning mechanism we also train models to synthesize images based on **semantic layouts** on OpenImages [49], and finetune on COCO [4], see Fig. 8. See Sec. D.3 for the quantitative evaluation and implementation details.

Lastly, following prior work [3, 15, 21, 23], we evaluate our best-performing **class-conditional** ImageNet models with $f \in \{4, 8\}$ from Sec. 4.1 in Tab. 3, Fig. 4 and Sec. D.4. Here we outperform the state of the art diffusion model ADM [15] while significantly reducing computational requirements and parameter count, *cf*. Tab 18.

4.3.2 Convolutional Sampling Beyond 256^2

By concatenating spatially aligned conditioning information to the input of ϵ_θ, *LDMs* can serve as efficient general-purpose image-to-image translation models. We use this to train models for semantic synthesis, super-resolution (Sec. 4.4) and inpainting (Sec. 4.5). For semantic synthesis, we use images of landscapes paired with semantic maps [23, 61] and concatenate downsampled versions of the semantic maps with the latent image representation of a $f = 4$ model (VQ-reg., see Tab. 8). We train on an input resolution of 256^2 (crops from 384^2) but find that our model generalizes to larger resolutions and can generate images up to the megapixel regime when evaluated in a convolutional manner (see Fig. 9). We exploit this behavior to also apply the super-resolution models in Sec. 4.4 and the inpainting models in Sec. 4.5 to generate large images between 512^2 and 1024^2. For this application, the signal-to-noise ratio (induced by the scale of the latent space) significantly affects the results. In Sec. D.1 we illustrate this when learning an LDM on (i) the latent space as provided by a $f = 4$ model (KL-reg., see Tab. 8), and (ii) a rescaled version, scaled by the component-wise standard deviation.

The latter, in combination with classifier-free guidance [32], also enables the direct synthesis of $> 256^2$ images for the text-conditional *LDM-KL-8-G* as in Fig. 13.

Method	FID↓	IS↑	Precision↑	Recall↑	N_{params}	
BigGan-deep [3]	6.95	203.6$_{\pm 2.6}$	**0.87**	0.28	340M	-
ADM [15]	10.94	100.98	0.69	**0.63**	554M	250 DDIM steps
ADM-G [15]	4.59	186.7	0.82	0.52	608M	250 DDIM steps
LDM-4 (ours)	10.56	103.49$_{\pm 1.24}$	0.71	0.62	400M	250 DDIM steps
LDM-4-G (ours)	**3.60**	**247.67**$_{\pm 5.59}$	**0.87**	0.48	400M	250 steps, c.f.g [32], $s = 1.5$

Table 3. Comparison of a class-conditional ImageNet *LDM* with recent state-of-the-art methods for class-conditional image generation on ImageNet [12]. A more detailed comparison with additional baselines can be found in D.4, Tab. 10 and F. *c.f.g.* denotes classifier-free guidance with a scale s as proposed in [32].

Figure 9. A *LDM* trained on 256^2 resolution can generalize to larger resolution (here: 512×1024) for spatially conditioned tasks such as semantic synthesis of landscape images. See Sec. 4.3.2.

4.4. Super-Resolution with Latent Diffusion

LDMs can be efficiently trained for super-resolution by diretly conditioning on low-resolution images via concatenation (*cf*. Sec. 3.3). In a first experiment, we follow SR3

Figure 10. ImageNet 64→256 super-resolution on ImageNet-Val. *LDM-SR* has advantages at rendering realistic textures but SR3 can synthesize more coherent fine structures. See appendix for additional samples and cropouts. SR3 results from [72].

[72] and fix the image degradation to a bicubic interpolation with $4\times$-downsampling and train on ImageNet following SR3's data processing pipeline. We use the $f = 4$ autoencoding model pretrained on OpenImages (VQ-reg., *cf*. Tab. 8) and concatenate the low-resolution conditioning y and the inputs to the UNet, *i.e*. τ_θ is the identity. Our qualitative and quantitative results (see Fig. 10 and Tab. 5) show competitive performance and LDM-SR outperforms SR3 in FID while SR3 has a better IS. A simple image regression model achieves the highest PSNR and SSIM scores; however these metrics do not align well with human perception [106] and favor blurriness over imperfectly aligned high frequency details [72]. Further, we conduct a user study comparing the pixel-baseline with LDM-SR. We follow SR3 [72] where human subjects were shown a low-res image in between two high-res images and asked for preference. The results in Tab. 4 affirm the good performance of LDM-SR. PSNR and SSIM can be pushed by using a post-hoc guiding mechanism [15] and we implement this *image-based guider* via a perceptual loss, see Sec. D.6.

	SR on ImageNet		Inpainting on Places	
User Study	Pixel-DM ($f1$)	*LDM-4*	LAMA [88]	*LDM-4*
Task 1: Preference vs GT ↑	16.0%	**30.4%**	13.6%	**21.0%**
Task 2: Preference Score ↑	29.4%	**70.6%**	31.9%	**68.1%**

Table 4. Task 1: Subjects were shown ground truth and generated image and asked for preference. Task 2: Subjects had to decide between two generated images. More details in E.3.6

Since the bicubic degradation process does not generalize well to images which do not follow this pre-processing, we also train a generic model, *LDM-BSR*, by using more diverse degradation. The results are shown in Sec. D.6.1.

Method	FID ↓	IS ↑	PSNR ↑	SSIM ↑	N_{params}	[samples/s]$^{(*)}$
Image Regression [72]	15.2	121.1	**27.9**	**0.801**	625M	N/A
SR3 [72]	5.2	**180.1**	26.4	0.762	625M	N/A
LDM-4 (ours, 100 steps)	2.8†/4.8‡	166.3	24.4$\pm$3.8	0.69$\pm$0.14	**169M**	4.62
emphLDM-4 (ours, big, 100 steps)	**2.4**†/**4.3**‡	174.9	24.7$\pm$4.1	0.71$\pm$0.15	552M	4.5
LDM-4 (ours, 50 steps, guiding)	4.4†/6.4‡	153.7	25.8$\pm$3.7	0.74$\pm$0.12	184M	0.38

Table 5. $\times 4$ upscaling results on ImageNet-Val. (256^2); †: FID features computed on validation split, ‡: FID features computed on train split; *: Assessed on a NVIDIA A100

Model (*reg.*-type)	train throughput samples/sec.	sampling throughput† @256	@512	train+val hours/epoch	FID@2k epoch 6
LDM-1 (no first stage)	0.11	0.26	0.07	20.66	24.74
LDM-4 (*KL*, w/ attn)	0.32	0.97	0.34	7.66	15.21
LDM-4 (*VQ*, w/ attn)	0.33	0.97	0.34	7.04	14.99
LDM-4 (*VQ*, w/o attn)	0.35	0.99	0.36	6.66	15.95

Table 6. Assessing inpainting efficiency. †: Deviations from Fig. 7 due to varying GPU settings/batch sizes *cf*. the supplement.

4.5. Inpainting with Latent Diffusion

Inpainting is the task of filling masked regions of an image with new content either because parts of the image are are corrupted or to replace existing but undesired content within the image. We evaluate how our general approach for conditional image generation compares to more specialized, state-of-the-art approaches for this task. Our evaluation follows the protocol of LaMa [88], a recent inpainting model that introduces a specialized architecture relying on Fast Fourier Convolutions [8]. The exact training & evaluation protocol on Places [108] is described in Sec. E.2.2.

We first analyze the effect of different design choices for the first stage. In particular, we compare the inpainting efficiency of *LDM-1* (*i.e*. a pixel-based conditional DM) with *LDM-4*, for both *KL* and *VQ* regularizations, as well as *VQ-LDM-4* without any attention in the first stage (see Tab. 8), where the latter reduces GPU memory for decoding at high resolutions. For comparability, we fix the number of parameters for all models. Tab. 6 reports the training and sampling throughput at resolution 256^2 and 512^2, the total training time in hours per epoch and the FID score on the validation split after six epochs. Overall, we observe a speed-up of at least $2.7\times$ between pixel- and latent-based diffusion models while improving FID scores by a factor of at least $1.6\times$.

The comparison with other inpainting approaches in Tab. 7 shows that our model with attention improves the overall image quality as measured by FID over that of [88]. LPIPS between the unmasked images and our samples is slightly higher than that of [88]. We attribute this to [88] only producing a single result which tends to recover more of an average image compared to the diverse results produced by our LDM *cf*. Fig. 21. Additionally in a user study (Tab. 4) human subjects favor our results over those of [88].

Based on these initial results, we also trained a larger diffusion model (*big* in Tab. 7) in the latent space of the *VQ*-regularized first stage without attention. Following [15], the UNet of this diffusion model uses attention layers on three levels of its feature hierarchy, the BigGAN [3] residual block for up- and downsampling and has 387M parameters

Figure 11. Qualitative results on object removal with our *big, w/ ft* inpainting model. For more results, see Fig. 22.

instead of 215M. After training, we noticed a discrepancy in the quality of samples produced at resolutions 256^2 and 512^2, which we hypothesize to be caused by the additional attention modules. However, fine-tuning the model for half an epoch at resolution 512^2 allows the model to adjust to the new feature statistics and sets a new state of the art FID on image inpainting (*big, w/o attn, w/ ft* in Tab. 7, Fig. 11.).

5. Limitations & Societal Impact

Limitations While LDMs significantly reduce computational requirements compared to pixel-based approaches, their sequential sampling process is still slower than that of GANs. Moreover, the use of LDMs can be questionable when high precision is required: although the loss of image quality is very small in our $f = 4$ autoencoding models (see Fig. 1), their reconstruction capability can become a bottleneck for tasks that require fine-grained accuracy in pixel space. We assume that our superresolution models (Sec. 4.4) are already somewhat limited in this respect.

Societal Impact Generative models for media like imagery are a double-edged sword: On the one hand, they enable various creative applications, and in particular approaches like ours that reduce the cost of training and inference have the potential to facilitate access to this technology and democratize its exploration. On the other hand, it also means that it becomes easier to create and disseminate manipulated data or spread misinformation and spam. In particular, the deliberate manipulation of images ("deep fakes") is a common problem in this context, and women in particular are disproportionately affected by it [13, 24].

Generative models can also reveal their training data [5, 90], which is of great concern when the data contain sensitive or personal information and were collected without explicit consent. However, the extent to which this also applies to DMs of images is not yet fully understood.

Finally, deep learning modules tend to reproduce or exacerbate biases that are already present in the data [22, 38, 91]. While diffusion models achieve better coverage of the data distribution than *e.g.* GAN-based approaches, the extent to which our two-stage approach that combines adversarial training and a likelihood-based objective misrepresents the data remains an important research question.

For a more general, detailed discussion of the ethical considerations of deep generative models, see *e.g.* [13].

	40-50% masked		All samples	
Method	FID ↓	LPIPS ↓	FID ↓	LPIPS ↓
LDM-4 (ours, big, w/ ft)	**9.39**	0.246± 0.042	**1.50**	0.137± 0.080
LDM-4 (ours, big, w/o ft)	12.89	0.257± 0.047	2.40	0.142± 0.085
LDM-4 (ours, w/ attn)	11.87	0.257± 0.042	2.15	0.144± 0.084
LDM-4 (ours, w/o attn)	12.60	0.259± 0.041	2.37	0.145± 0.084
LaMa [88]†	12.31	**0.243**± 0.038	2.23	**0.134**± 0.080
LaMa [88]	12.0	**0.24**	2.21	0.14
CoModGAN [107]	10.4	0.26	1.82	0.15
RegionWise [52]	21.3	0.27	4.75	0.15
DeepFill v2 [104]	22.1	0.28	5.20	0.16
EdgeConnect [58]	30.5	0.28	8.37	0.16

Table 7. Comparison of inpainting performance on 30k crops of size 512×512 from test images of Places [108]. The column *40-50%* reports metrics computed over hard examples where 40-50% of the image region have to be inpainted. †recomputed on our test set, since the original test set used in [88] was not available.

6. Conclusion

We have presented latent diffusion models, a simple and efficient way to significantly improve both the training and sampling efficiency of denoising diffusion models without degrading their quality. Based on this and our cross-attention conditioning mechanism, our experiments could demonstrate favorable results compared to state-of-the-art methods across a wide range of conditional image synthesis tasks without task-specific architectures.

This work has been supported by the German Federal Ministry for Economic Affairs and Energy within the project 'KI-Absicherung - Safe AI for automated driving' and by the German Research Foundation (DFG) project 421703927.