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.

高分辨率图像合成与潜在扩散模型

Robin Rombach[1] * Andreas Blattmann[1] * Dominik Lorenz[1] Patrick Esser Björn Ommer[1]

[1]Ludwig Maximilian University of Munich & IWR, Heidelberg University, Germany Runway ML https://github.com/CompVis/latent-diffusion

摘要

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. 引言

图像合成是计算机视觉领域中近期发展最为惊人的方向之一，同时也是计算需求最为庞大的领域之一。尤其是复杂自然场景的高分辨率合成，目前主要依赖于基于似然的模型规模化扩展，这些模型在自回归（AR）变换器中可能包含数十亿参数[66,67]。相比之下，生成对抗网络（GANs）[3,27,40]所展现的优异成果，大多局限于数据变异性相对有限的情况，因为其对抗性学习过程难以扩展到对复杂多模态分布的建模。最近，基于层级去噪自编码器构建的扩散模型[82]已展现出令人瞩目的

图1. 通过降低下采样强度提升可达到质量的上限。由于扩散模型为空间数据提供了优异的归纳偏置，我们无需像相关潜空间生成模型那样进行剧烈的空间下采样，仍能通过合适的自编码模型大幅降低数据维度（详见第3节）。图像来自DIV2K[1]验证集，评估分辨率为512^2像素。空间下采样因子记为f。重建FID[29]与PSNR在ImageNet-val[12]上计算；另见表8。

在图像合成[30,85]及其他领域[7,45,48,57]中取得了显著成果，并在类别条件图像合成[15,31]与超分辨率[72]任务上定义了当前最优性能。此外，即使是无条件的扩散模型也能直接应用于图像修复、着色[85]或笔触合成[53]等任务，这与其他类型的生成模型[19,46,69]形成鲜明对比。作为基于似然的模型，它们不会出现GAN中常见的模式崩溃和训练不稳定问题，并且通过充分利用参数共享机制，能够以远少于自回归模型[67]所需的数十亿参数规模，对高度复杂的自然图像分布进行建模。

高分辨率图像合成的民主化 扩散模型属于基于似然的模型类别，其模式覆盖特性使它们倾向于将过多的容量（以及计算资源）用于建模数据中难以察觉的细节[16, 73]。尽管重加权变分目标[30]旨在通过欠采样初始去噪步骤来解决这一问题，但扩散模型仍然对计算资源要求很高，因为训练和评估此类模型需要在RGB图像的高维空间中进行重复的函数评估（和梯度计算）。例如，训练最强大的扩散模型通常需要数百个GPU天（在[15]中为{v*}150-1000个V100天），并且在输入空间的噪声版本上进行重复评估也使得推理成本高昂。

*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.

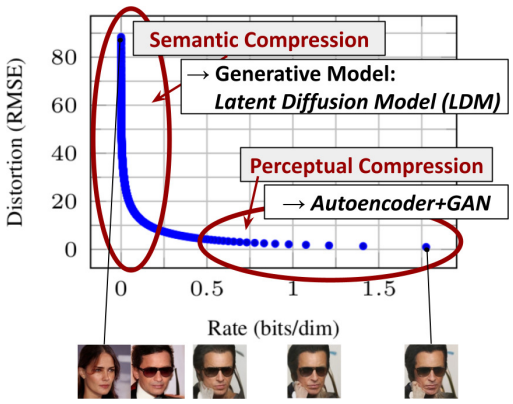

Figure 2. Illustrating perceptual and semantic compression: Most bits of a digital image correspond to imperceptible details. While DMs allow to suppress this semantically meaningless information by minimizing the responsible loss term, gradients (during training) and the neural network backbone (training 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-

因此，在单个A100 GPU上生成5万个样本大约需要5天时间[15]。这对研究界和普通用户产生了两个影响：首先，训练这样一个模型需要巨大的计算资源，只有该领域的一小部分人能够获得，并且会留下巨大的碳足迹[65, 86]。其次，评估一个已经训练好的模型在时间和内存上同样昂贵，因为相同的模型架构必须按顺序运行大量步骤（在[15]中为{v*}25至1000步）。

为了提升这一强大模型类别的可及性，同时降低其显著的资源消耗，我们需要一种能够减少训练和采样计算复杂度的方法。因此，在不损害其性能的前提下降低扩散模型的计算需求，是增强其可及性的关键。

前往潜在空间 我们的方法始于分析像素空间中已训练好的扩散模型：图2展示了一个已训练模型的率失真权衡。与任何基于似然的模型一样，学习过程大致可分为两个阶段：首先是*perceptualcompression*阶段，该阶段会去除高频细节，但几乎不学习语义变化。在第二阶段，真正的生成模型学习数据的语义和概念构成(*semantic compression*)。因此，我们的目标是首先找到一个感知上等效但计算上更合适的空间，在该空间中训练扩散模型以实现高分辨率图像合成。

遵循常见做法[11, 23, 66, 67, 96]，我们将训练分为两个不同的阶段：首先，我们训练一个自动编码器，它提供一个与数据空间感知等效的低维（从而高效）表示空间。重要的是，与先前工作[23,66]相比，我们无需依赖过度的空间压缩，因为我们在学习到的潜在空间中训练扩散模型，该空间在空间维度方面展现出更好的扩展特性。降低的复杂度还使得通过单次网络前向传递即可从潜在空间高效生成图像。我们将得到的模型类别称为*Latent Diffusion Models* (潜在扩散模型)。

这种方法的一个显著优势在于，我们只需训练一次通用自编码阶段，之后便可将其重复用于多个扩散模型训练，或探索可能完全不同的任务[81]。这使得我们能够高效探索大量用于各类图像到图像及文本到图像任务的扩散模型。针对文本到图像任务，我们设计了一种将变换器连接到扩散模型UNet主干的结构[71]，该结构支持任意类型的基于令牌的条件机制，详见第3.3节。

总之，我们的工作做出了以下贡献：

(i) 与纯基于Transformer的方法[23, 66]相比，我们的方法能更优雅地扩展到更高维数据，从而能够：(a) 在压缩层级上工作，提供比以往工作更忠实、更细致的重建结果（见图1）；(b) 高效地

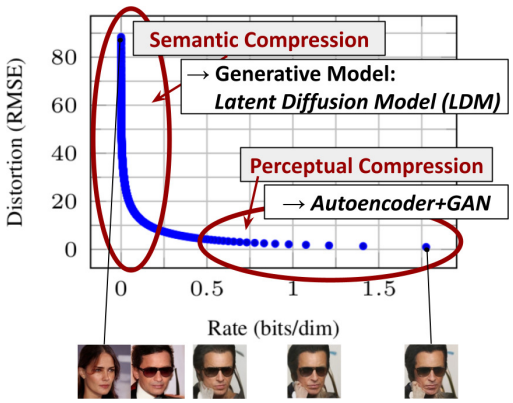

图2. 感知压缩与语义压缩示意图：数字图像的大部分比特对应不可察觉的细节。虽然扩散模型可通过最小化相应损失项来抑制这类语义无意义信息，但梯度（训练期间）和神经网络主干（训练与推理阶段）仍需在所有像素上进行计算，导致冗余运算及不必要的昂贵优化与推理开销。我们提出{v*}作为一种高效生成模型，并引入独立的轻量压缩阶段，仅消除不可感知的细节。数据与图像来源自[30]。

应用于百万像素图像的高分辨率合成。

(ii) 我们在多个任务（无条件图像合成、修复、随机超分辨率）和数据集上实现了具有竞争力的性能，同时显著降低了计算成本。与基于像素的扩散方法相比，我们还显著降低了推理成本。

(iii) 与先前需要同时学习编码器/解码器架构和基于分数的先验的工作[93]不同，我们的方法无需在重构能力与生成能力之间进行精细权衡。这确保了极高保真度的重构，并且对隐空间的正则化需求极低。

(iv) 我们发现，对于超分辨率、修复和语义合成等密集条件任务，我们的模型可以以卷积方式应用，并渲染出~1024^2像素的大尺寸、一致性图像。

(v) 此外，我们设计了一种基于交叉注意力的通用条件调节机制，支持多模态训练。我们利用该机制训练了类别条件生成模型、文本到图像生成模型以及布局到图像生成模型。

(vi) 最后，我们在 https://github.com/CompVis/latent-diffusion 发布了预训练的潜在扩散和自编码模型，这些模型除了用于训练 DM [81] 外，还可能适用于各种任务。

2. 相关工作

图像生成的生成模型　图像的高维特性给生成模型带来了独特的挑战。生成对抗网络（GAN）[27] 能够高效采样具有良好感知质量的高分辨率图像 [3, 42]，但其训

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-

难以优化[2, 28, 54]，且难以捕捉完整的数据分布[55]。相比之下，基于似然的方法强调良好的密度估计，这使得优化过程更加稳定。变分自编码器（VAE）[46]和基于流的模型[18, 19]能够高效合成高分辨率图像[9, 44, 92]，但样本质量仍不及生成对抗网络（GAN）。自回归模型（ARM）[6, 10, 94, 95]在密度估计方面表现优异，但计算密集的架构[97]和顺序采样过程限制了其只能处理低分辨率图像。由于基于像素的图像表示包含难以察觉的高频细节[16,73]，最大似然训练会耗费不成比例的计算能力来建模这些细节，导致训练时间漫长。为了扩展到更高分辨率，一些两阶段方法[23, 67,101,103]使用自回归模型来建模压缩的潜在图像空间，而非原始像素。

最近，扩散概率模型（DM）[82]在密度估计[45]和样本质量[15]方面均取得了最先进的结果。这些模型的生成能力源于其底层神经骨干网络采用UNet[15, 30, 71, 85]实现时，能自然契合类图像数据的归纳偏置。通常，使用加权目标函数[30]进行训练时可获得最佳合成质量。此时，DM相当于一个有损压缩器，允许在图像质量和压缩能力之间进行权衡。然而，在像素空间评估和优化这些模型存在推理速度慢、训练成本极高的缺点。虽然前者可通过先进采样策略[47, 75, 84]和分层方法[31, 93]部分缓解，但高分辨率图像数据的训练始终需要计算昂贵的梯度。我们提出的{v*}通过作用于低维压缩潜在空间，同时解决了这两个缺陷。这使得训练计算成本更低，并加速推理过程，且合成质量几乎不受影响（见图1）。

两阶段图像合成 为了弥补单一生成方法的不足，大量研究[11, 23, 67, 70, 101, 103]通过两阶段方法将不同技术的优势结合到更高效、性能更强的模型中。VQ-VAE[67, 101]利用自回归模型在离散化潜空间上学习富有表现力的先验分布。[66]通过联合学习离散化图像与文本表示的分布，将该方法扩展到文本到图像生成领域。更广义地说，[70]采用条件可逆网络实现跨领域潜空间的通用转换。与VQ-VAE不同，VQGAN[23, 103]在第一阶段采用对抗性与感知性目标，使自回归变换器能处理更大尺寸图像。然而，为适应ARM训练所需的高压缩率（这会引入数十亿可训练参数[23, 66]），限制了此类方法的整体性能。

方法以及较低的压缩率是以高计算成本为代价的[23, 66]。我们的工作避免了这种权衡，因为我们提出的*LDMs*由于其卷积骨干网络，能够更平缓地适应更高维度的潜在空间。因此，我们可以自由选择压缩级别，在确保高保真重建的同时（见图1），既能学习到强大的第一阶段，又不会给生成扩散模型留下过多的感知压缩负担。

尽管存在联合[93]或分别[80]学习编码/解码模型与基于分数的先验的方法，但前者仍需在重建能力和生成能力之间进行困难的权衡[11]，且性能不及我们的方法（第4节）；而后者则专注于高度结构化的图像（如人脸）。

3. 方法

为了降低训练扩散模型进行高分辨率图像合成的计算需求，我们注意到，尽管扩散模型允许通过欠采样相应的损失项来忽略感知上不相关的细节[30]，但它们仍然需要在像素空间进行昂贵的函数评估，这导致计算时间和能源资源的巨大消耗。

我们提出通过明确分离压缩学习阶段与生成学习阶段来规避这一缺陷（见图2）。为此，我们采用自编码模型，该模型学习一个在感知上与图像空间等效、但计算复杂度显著降低的空间。

这种方法具有以下几个优势：(i) 通过离开高维图像空间，我们获得的扩散模型在计算上更为高效，因为采样是在低维空间中进行的。(ii) 我们利用了扩散模型从其UNet架构[71]继承而来的归纳偏置，这使得它们对具有空间结构的数据特别有效，从而减少了对以往方法[23, 66]所需的那种会降低质量的激进压缩级别的需求。(iii) 最后，我们获得了通用的压缩模型，其潜在空间可用于训练多个生成模型，并且还可用于其他下游应用，例如单图像CLIP引导合成[25]。

3.1. 感知图像压缩

我们的感知压缩模型基于先前的工作[23]，由一个自编码器构成，该自编码器通过结合感知损失[106]和基于图像块的对抗目标[20, 23, 103]进行训练。这通过强制局部真实性，确保重建结果被限制在图像流形内，并避免了仅依赖像素空间损失（如L_2或L_1目标）所引入的模糊性。

更准确地说，给定RGB空间中的图像$x \in \mathbb{R}^{H\times W\times 3}$，编码器$\mathcal{E}$将$x$编码为潜在表示——

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

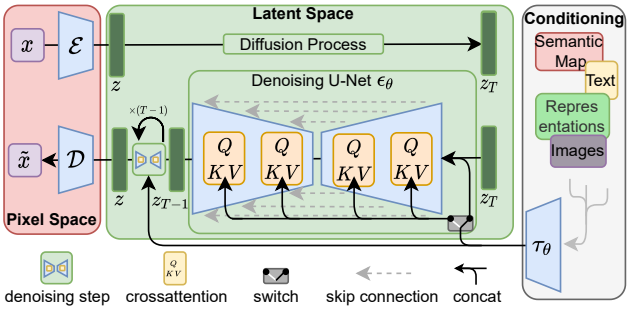

Figure 3. We condition LDMs either via concatenation or by a more general cross-attention mechanism. See Sec. 3.3

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

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 $W_V^{(i)} \in$

编码器 $z = \mathcal{E}(x)$ 将图像压缩为潜在表示，解码器 $\mathcal{D}$ 从潜在表示重建图像，得到 $\tilde{x} = \mathcal{D}(z) = \mathcal{D}(\mathcal{E}(x))$，其中 $z \in \mathbb{R}^{h \times w \times c}$。重要的是，编码器 *downsamples* 将图像下采样 $f = H/h = W/w$ 倍，我们研究了不同的下采样因子 $f = 2^m$，其中 $m \in \mathbb{N}$。

为避免潜在空间方差任意增大，我们尝试了两种不同的正则化方法。第一种变体 *KL-reg.* 对学习到的潜在表示施加轻微 KL 惩罚，使其趋近标准正态分布，类似于 VAE [46, 69]；而 *VQ-reg.* 则在解码器中使用了向量量化层 [96]。该模型可理解为 VQGAN [23]，但量化层被解码器吸收。由于我们后续的扩散模型专为二维结构的潜在空间 $z = \mathcal{E}(x)$ 设计，因此可采用相对温和的压缩率并获得出色的重建效果。这与先前研究 [23, 66] 形成对比——那些方法依赖学习空间 z 的任意一维排序进行自回归分布建模，从而忽略了 z 的固有结构。因此，我们的压缩模型能更好地保留 x 的细节（见表 8）。完整目标函数与训练细节详见附录。

3.2. 潜在扩散模型

扩散模型[82]是一种概率模型，旨在通过逐步去噪一个正态分布变量来学习数据分布$p(x)$，这相当于学习一个长度为T的固定马尔可夫链的逆向过程。在图像合成领域，最成功的模型[15,30,72]依赖于$p(x)$的变分下界的一个重新加权变体，这与去噪分数匹配[85]的原理一致。这些模型可以解释为一系列等权重的去噪自编码器$\epsilon_\theta(x_t, t)$；$t = 1 \dots T$，它们被训练用于预测其输入x_t的去噪变体，其中x_t是输入x的噪声版本。相应的目标函数可简化为（B节）

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

其中t从$\{1, \dots, T\}$中均匀采样。
潜在表示的生成建模 通过我们训练好的感知压缩模型（包括$\mathcal{E}$和$\mathcal{D}$），我们现在能够访问一个高效、低维的潜在空间，其中高频且难以察觉的细节已被抽象化。与高维像素空间相比，该空间更适合基于似然的生成模型，因为它们现在可以（i）专注于数据中重要的语义信息，并且（ii）在更低维度、计算效率显著更高的空间中进行训练。

与先前依赖高度压缩的离散潜在空间中自回归、基于注意力的Transformer模型的研究[23, 66, 103]不同，我们能够利用本模型所提供的图像特定归纳偏置。

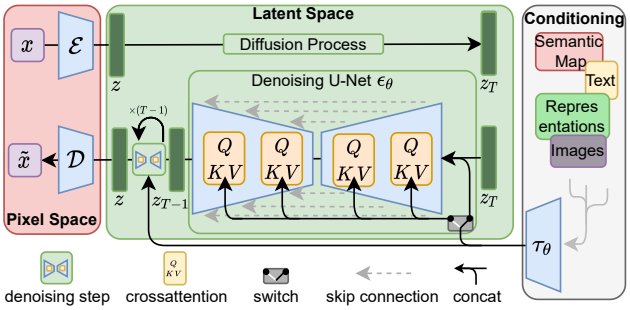

图3. 我们通过拼接或更通用的交叉注意力机制对LDMs进行条件控制。详见第3.3节

包括能够主要从2D卷积层构建底层UNet，并通过重新加权边界进一步将目标聚焦于感知上最相关的比特，该边界现在表述为

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

我们模型的神经主干 $\epsilon_\theta(\circ, t)$ 实现为一个时间条件UNet [71]。由于前向过程是固定的，z_t 在训练期间可以从 $\mathcal{E}$ 高效获取，且来自 $p(z)$ 的样本只需通过 $\mathcal{D}$ 单次前传即可解码到图像空间。

3.3. 条件机制

与其他类型的生成模型[56, 83]类似，扩散模型原则上能够建模形式为$p(z|y)$的条件分布。这可以通过条件去噪自编码器$\epsilon_\theta(z_t, t, y)$实现，并为通过文本[68]、语义图[33, 61]或其他图像到图像翻译任务[34]等输入y来控制合成过程铺平了道路。

然而，在图像合成领域，将扩散模型的生成能力与类别标签[15]或输入图像的模糊变体[72]之外的其他条件类型相结合，至今仍是一个研究不足的领域。

我们通过在其底层UNet骨干网络中引入交叉注意力机制[97]，将扩散模型转变为更灵活的条件图像生成器，该机制对于学习基于注意力的多模态输入模型非常有效[35,36]。为预处理来自不同模态（如语言提示）的y，我们引入一个领域专用编码器τ_θ，它将y投影到中间表示$\tau_\theta(y) \in \mathbb{R}^{M \times d_\tau}$，随后通过交叉注意力层将其映射到UNet的中间层，该层实现为Attention$(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V$，其中

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

这里，$\varphi_i(z_t) \in \mathbb{R}^{N \times d_\epsilon^i}$ 表示实现 ϵ_θ 和 $W_V^{(i)} \in$ 的 UNet 的（扁平化）中间表示。

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

图4. 在CelebAHQ [39]、FFHQ [41]、LSUN-Churches [102]、LSUN-Bedrooms [102] 和类别条件ImageNet [12] 上训练的 *LDMs* 样本，分辨率均为256 × 256。放大查看效果更佳。更多样本见 *cf* 补充材料。

$\mathbb{R}^{d\times d_\epsilon^i}$、$W_Q^{(i)} \in \mathbb{R}^{d\times d_\tau}$ 和 $W_K^{(i)} \in \mathbb{R}^{d\times d_\tau}$ 是可学习的投影矩阵 [36, 97]。视觉描述见图 3。

基于图像条件配对，我们随后通过以下方式学习条件LDM：

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

其中，τ_θ和ϵ_θ均通过公式3进行联合优化。这种条件机制具有灵活性，因为τ_θ可通过领域特定专家*e.g*进行参数化处理（未屏蔽的变换器[97]在y为文本提示时适用，参见第4.3.1节）。

4. 实验

LDMs 提供了灵活且计算上易于处理的基于扩散的各种图像模态合成方法，我们将在下文中通过实验展示这一点。然而，首先我们分析了我们的模型与基于像素的扩散模型在训练和推理方面的优势。有趣的是，我们发现，在*VQ*正则化的潜在空间中训练的*LDMs*有时能获得更好的样本质量，尽管*VQ*正则化的第一阶段模型的重建能力略逊于其连续对应模型*cf*。表8展示了第一阶段正则化方案对*LDM*训练效果及其向分辨率$>256^2$泛化能力的视觉比较，详见附录D.1。在E.2中，我们列出了本节所有结果在架构、实现、训练和评估方面的详细信息。

4.1. 关于感知压缩的权衡

本节分析了我们具有不同下采样因子 $f \in \{1, 2, 4, 8, 16, 32\}$ (（简记为 *LDM-f*）的LDMs行为，其中 *LDM-1* 对应基于像素的DMs)。为获得可比较的测试环境，我们在本节所有实验中固定计算资源为单张NVIDIA A100，并以相同的训练步数和参数量训练所有模型。

表8展示了用于*LDMs*比较的第一阶段模型的超参数和重建性能。

在本节中进行了比较。图6展示了在ImageNet[12]数据集上，经过200万步训练的类别条件模型，其样本质量随训练进程的变化情况。我们发现：i) 较小的降采样因子*LDM-{1,2}*会导致训练进展缓慢；而ii) 过大的*f*值则会在相对较少的训练步数后导致生成保真度停滞不前。回顾上文分析（图1和图2），我们将此归因于：i) 将大部分感知压缩任务留给了扩散模型；ii) 第一阶段压缩过强导致信息丢失，从而限制了可达到的质量。*LDM-{4-16}*在效率与感知保真度之间取得了良好平衡，这体现在经过200万步训练后，基于像素的扩散模型(*LDM-1*)与*LDM-8*之间存在高达38的显著FID[29]差距。

在图7中，我们比较了在CelebA-HQ [39]和ImageNet上训练的模型，使用DDIM采样器[84]在不同去噪步数下的采样速度，并将其与FID分数[29]进行对比。*LDM-{4-8}*在感知压缩与概念压缩比例不合适的模型中表现更优。特别是与基于像素的*LDM-1*相比，它们在显著提高样本吞吐量的同时，获得了更低的FID分数。对于复杂数据集（如ImageNet），需要降低压缩率以避免质量损失。总之，*LDM-4*和*-8*为实现高质量合成结果提供了最佳条件。

4.2. 基于潜在扩散的图像生成

我们在CelebA-HQ [39]、FFHQ [41]、LSUN-Churches和-Bedrooms [102]上训练了256^2图像的无条件模型，并通过i) FID [29]和ii) Precision-and-Recall [50]评估了i)样本质量与ii)模型对数据流形的覆盖能力。表1总结了我们的结果。在CelebA-HQ上，我们实现了5.11的最新最优FID，超越了之前的基于似然的模型以及GANs。我们还超越了LSGM [93]（其中潜在扩散模型与第一阶段联合训练）。相比之下，我们在固定空间中训练扩散模型。

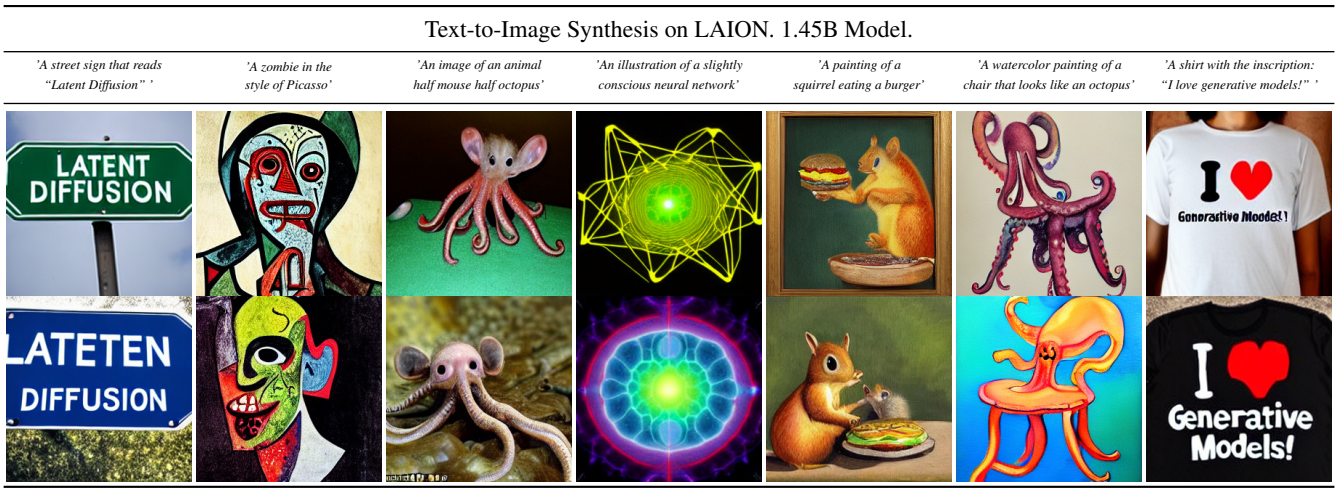

Figure 5. Samples for user-defined text prompts from our model for text-to-image synthesis, *LDM-8 (KL)*, which was trained on the LAION [78] database. Samples generated with 200 DDIM steps and $\eta = 1.0$. We use unconditional guidance [32] with $s = 10.0$.

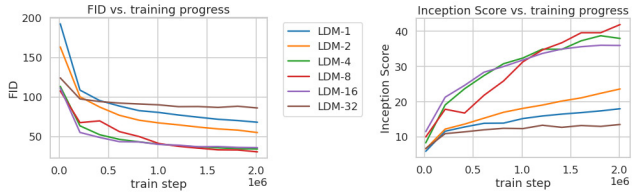

Figure 6. Analyzing the training of class-conditional *LDMs* with different downsampling factors f over 2M train steps on the ImageNet dataset. Pixel-based *LDM-1* requires substantially larger train times compared to models with larger downsampling factors (*LDM-*$\{$*4-16*$\}$). Too much perceptual 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 $\kappa = 0$.

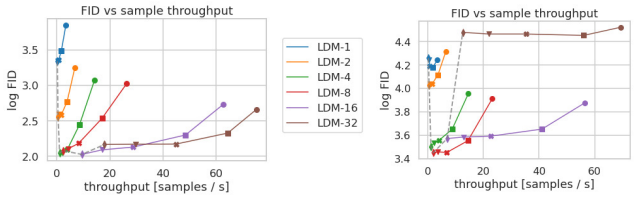

Figure 7. Comparing *LDMs* with varying compression on the CelebA-HQ (left) and ImageNet (right) datasets. Different markers indicate $\{10, 20, 50, 100, 200\}$ sampling steps using DDIM, from right to left along each line. The dashed line shows the FID scores for 200 steps, indicating the strong performance 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.

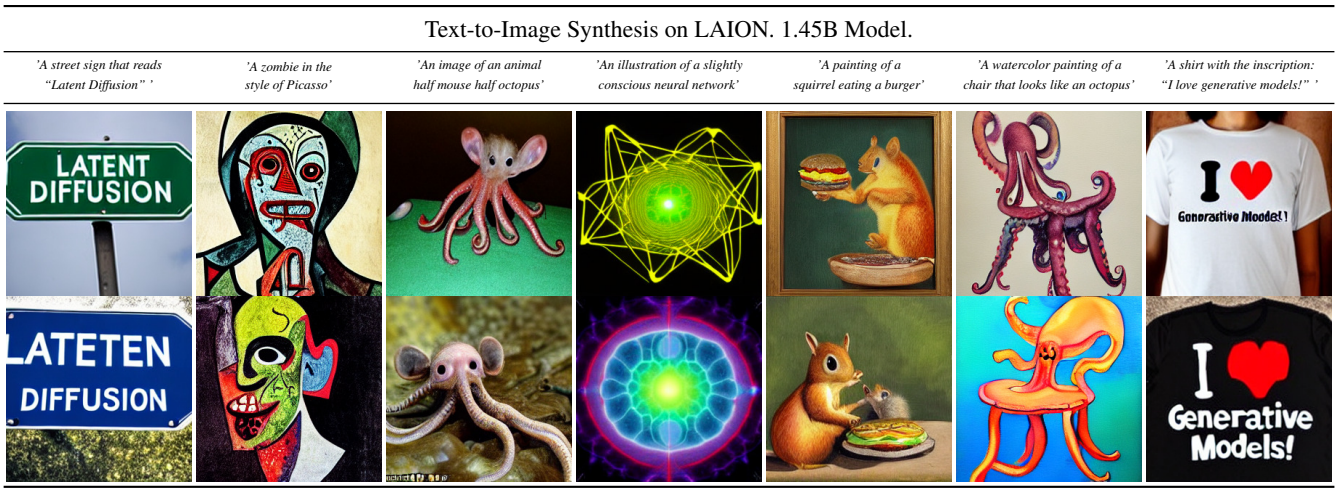

图5. 我们模型在文本到图像合成中用户自定义文本提示的样本，*LDM-8 (KL)*，该模型在LAION [78]数据库上训练。样本使用200步DDIM和$\eta = 1.0$生成。我们采用无条件引导[32]方法，参数为$s = 10.0$。

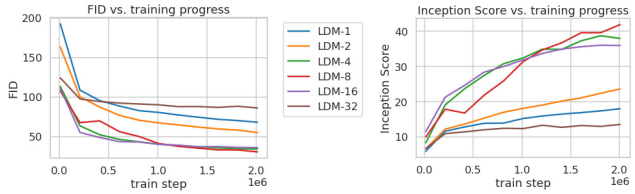

图6. 在ImageNet数据集上分析经过200万训练步数、采用不同下采样因子f的类别条件*LDMs*训练过程。基于像素的*LDM-1*相比具有更大下采样因子的模型(*LDM-*{*4-16*})需要显著更长的训练时间。如*LDM-32*中过度的感知压缩会限制整体样本质量。所有模型均在单张NVIDIA A100显卡上以相同计算预算训练。结果通过100步DDIM采样[84]及$\kappa = 0$获得。

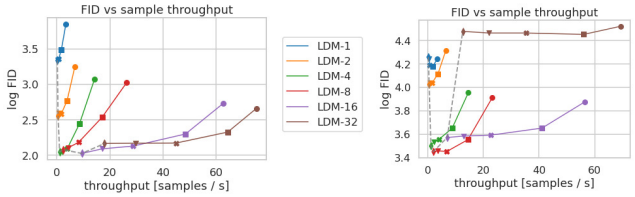

图7. 在CelebA-HQ（左）和ImageNet（右）数据集上比较不同压缩下的*LDMs*。不同标记表示使用DDIM的{10, 20, 50, 100, 200}采样步数，沿每条线从右到左排列。虚线显示200步的FID分数，表明*LDM-*{*4-8*}的优异性能。FID分数基于5000个样本评估。所有模型均在A100上训练了500k（CelebA）/ 2M（ImageNet）步。

并避免了在权衡重构质量与学习潜在空间先验之间的困难，参见图1-2。

在除LSUN-Bedrooms数据集外的所有任务中，我们的表现均优于先前的扩散方法；而在LSUN-Bedrooms数据集上，尽管我们仅使用一半参数且训练资源需求减少四倍（见附录E.3.5），其得分仍与ADM [15] 接近。

CelebA-HQ 256 × 256

Method	FID ↓	Prec. ↑	Recall ↑
DC-VAE [63]	15.8	-	-
VQGAN+T. [23] (k=400)	10.2	-	-
PGGAN [39]	8.0	-	-
LSGM [93]	7.22	-	-
UDM [43]	7.16	-	-
LDM-4 (ours, 500-s†)	**5.11**	0.72	0.49

FFHQ 256 × 256

Method	FID ↓	Prec. ↑	Recall ↑
ImageBART [21]	9.57	-	-
U-Net GAN (+aug) [77]	10.9 (7.6)	-	-
UDM [43]	5.54	-	-
StyleGAN [41]	4.16	0.71	0.46
ProjectedGAN [76]	**3.08**	0.65	0.46
LDM-4 (ours, 200-s)	4.98	**0.73**	**0.50**

LSUN-Churches 256 × 256

Method	FID ↓	Prec. ↑	Recall ↑
DDPM [30]	7.89	-	-
ImageBART [21]	7.32	-	-
PGGAN [39]	6.42	-	-
StyleGAN [41]	4.21	-	-
StyleGAN2 [42]	3.86	-	-
ProjectedGAN [76]	**1.59**	0.61	0.44
*LDM-8** (ours, 200-s)	4.02	**0.64**	**0.52**

LSUN-Bedrooms 256 × 256

Method	FID ↓	Prec. ↑	Recall ↑
ImageBART [21]	5.51	-	-
DDPM [30]	4.9	-	-
UDM [43]	4.57	-	-
StyleGAN [41]	2.35	0.59	0.48
ADM [15]	1.90	**0.66**	**0.51**
ProjectedGAN [76]	**1.52**	0.61	0.34
LDM-4 (ours, 200-s)	2.95	**0.66**	0.48

表1. 无条件图像合成的评估指标。CelebA-HQ结果复现自[43, 63, 100]，FFHQ结果复现自[42, 43]。†：N-s指使用DDIM[84]采样器进行N步采样。*：在*KL*正则化的潜在空间中训练。更多结果可参见补充材料。

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] $s = 1.5$

表2. 在256×256尺寸MS-COCO [51]数据集上对文本条件图像合成的评估：使用250步DDIM [84]采样时，我们的模型与最新的扩散模型[59]和自回归模型[26]方法性能相当，尽管使用的参数量显著更少。†/*：数据来自[109]/[26]

此外，*LDMs*在精确度和召回率上持续优于基于GAN的方法，从而证实了其基于模式覆盖的似然训练目标相较于对抗性方法的优势。在图4中，我们还展示了各数据集的定性结果。

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

图8. 在COCO数据集[4]上使用*LDM*进行的布局到图像合成，详见第4.3.1节。定量评估见补充材料D.3。

4.3. 条件潜在扩散

4.3.1 用于LDMs的Transformer编码器

通过将基于交叉注意力的条件机制引入LDMs，我们为扩散模型开启了先前未被探索的多种条件模态。在文本到图像建模任务中，我们训练了一个拥有14.5亿参数的*KL*正则化*LDM*模型，该模型以LAION-400M数据集[78]上的语言提示为条件。我们采用BERT分词器[14]并将τ_θ实现为Transformer架构[97]，以推断出通过（多头）交叉注意力映射到UNet中的潜代码（见第3.3节）。这种结合领域专家学习语言表征与视觉合成的方案，形成了一个强大的模型，能够很好地泛化到复杂且用户自定义的文本提示，*cf*。见图8和图5。在定量分析方面，我们遵循先前工作，在MS-COCO[51]验证集上评估文本到图像生成任务，我们的模型在强大的自回归方法[17,66]和基于GAN的方法[109]基础上实现了提升，*cf*。见表2。我们注意到，应用无分类器扩散引导[32]显著提升了样本质量，使得经过引导的*LDM-KL-8-G*在文本到图像合成任务中与最新的先进自回归模型[26]及扩散模型[59]性能相当，同时大幅减少了参数量。为进一步分析基于交叉注意力的条件机制的灵活性，我们还训练了基于OpenImages[49]语义布局合成图像的模型，并在COCO[4]上进行了微调，见图8。定量评估及实现细节参见附录D.3节。

最后，遵循先前的工作[3, 15, 21, 23]，我们在表3、图4和附录D.4中，使用第4.1节中的$f \in \{4,8\}$评估了我们性能最佳的类别条件ImageNet模型。在此，我们超越了最先进的扩散模型ADM[15]，同时显著降低了计算需求和参数数量*cf*。见表18。

4.3.2 超越256^2的卷积采样

通过将空间对齐的条件信息与ϵ_θ的输入连接起来，*LDMs*可以作为一种高效的通用

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], $s = 1.5$

表3. 在ImageNet数据集上，类别条件ImageNet *LDM*与近期最先进的类别条件图像生成方法的比较[12]。与更多基线的详细比较可参见D.4节表10及附录F。*c.f.g.*表示采用[32]提出的尺度为s的无分类器引导方法。

目的图像到图像转换模型。我们利用此技术训练模型进行语义合成、超分辨率（第4.4节）和修复（第4.5节）。对于语义合成，我们使用与语义地图配对的景观图像[23, 61]，并将语义地图的下采样版本与$f = 4$模型的潜在图像表示（VQ-reg.，见表8）进行拼接。我们在256^2 (裁剪自384^2)的输入分辨率上进行训练，但发现我们的模型能够泛化到更高分辨率，并且当以卷积方式评估时，可以生成高达百万像素级别的图像（见图9）。我们利用这一特性，将第4.4节中的超分辨率模型和第4.5节中的修复模型应用于生成512^2至1024^2之间的大尺寸图像。在此应用中，信噪比（由潜在空间的尺度引起）显著影响结果。在第D.1节中，我们通过以下两种方式说明了这一点：在（i）由$f = 4$模型提供的潜在空间（KL-reg.，见表8）上学习LDM，以及（ii）按分量标准差缩放的重新缩放版本上学习LDM。

后者与无分类器引导[32]相结合，还能直接合成$> 256^2$图像，用于文本条件*LDM-KL-8-G*，如图13所示。

图9. 在256^2分辨率上训练的*LDM*能够泛化至更高分辨率（此处为512×1024），适用于空间条件任务，如景观图像的语义合成。详见章节4.3.2。

4.4. 基于隐扩散模型的超分辨率

LDMs可以通过直接与低分辨率图像进行拼接（{v*}第3.3节）来实现高效的超分辨率训练。在首个实验中，我们遵循SR3的方法。

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

图10. ImageNet 64→256 在 ImageNet-Val 上的超分辨率结果。*LDM-SR* 在渲染真实纹理方面具有优势，但 SR3 能合成更连贯的精细结构。更多样本及局部裁剪图见附录。SR3 结果来自文献[72]。

[72] 并将图像退化固定为4×下采样的双三次插值，随后在ImageNet上按照SR3的数据处理流程进行训练。我们使用在OpenImages上预训练的$f = 4$自编码模型（VQ-reg., *cf*。表8），并将低分辨率条件y与输入连接到UNet，*i.e*。τ_θ是恒等映射。我们的定性和定量结果（见图10和表5）显示出具有竞争力的性能，LDM-SR在FID上优于SR3，而SR3的IS更好。一个简单的图像回归模型获得了最高的PSNR和SSIM分数；然而，这些指标与人类感知[106]的一致性不佳，并且更倾向于模糊而非未完美对齐的高频细节[72]。此外，我们进行了一项用户研究，比较像素基线方法与LDM-SR。我们遵循SR3[72]的方法，向人类受试者展示一张低分辨率图像和两张高分辨率图像，并询问其偏好。表4中的结果证实了LDM-SR的良好性能。PSNR和SSIM可以通过使用事后引导机制[15]来提升，我们通过感知损失实现了这一*image-based guider*，详见D.6节。

	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%**

表4. 任务1：向受试者展示真实图像与生成图像并要求选择偏好。任务2：受试者需在两幅生成图像间做出选择。更多细节见E.3.6节。

由于双三次退化过程对不遵循此预处理的图像泛化能力不佳，我们还通过使用更多样化的退化方式训练了一个通用模型*LDM-BSR*。结果展示在D.6.1节中。

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

表5. ×4 在ImageNet-Val上的放大结果。(256^2); †: FID特征在验证集上计算，‡: FID特征在训练集上计算；*: 在NVIDIA A100上评估

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

表6. 评估修复效率。†：由于GPU设置/批次大小不同导致的与图7的偏差*cf*。补充材料。

4.5. 基于潜在扩散模型的图像修复

图像修复的任务是向图像的掩码区域填充新内容，这通常是因为图像部分损坏，或是为了替换图像中现有但不理想的内容。我们评估了在此任务中，我们用于条件图像生成的通用方法与更专业化、最先进的方法相比表现如何。我们的评估遵循了LaMa[88]的协议，这是一个最近的图像修复模型，它引入了依赖快速傅里叶卷积[8]的专用架构。在Places[108]数据集上的具体训练与评估协议详见E.2.2节。

我们首先分析了第一阶段不同设计选择的效果。具体而言，我们比较了基于像素的条件扩散模型)与*LDM-4*在*KL*和*VQ*两种正则化方式下的修复效率，以及第一阶段不含注意力机制的*VQ-LDM-4*（见表8），后者可降低高分辨率解码时的GPU内存占用。为确保可比性，我们固定了所有模型的参数量。表6展示了在256^2和512^2分辨率下的训练与采样吞吐量、每轮训练总时长（小时）以及六轮训练后在验证集上的FID分数。总体而言，我们观察到基于潜空间的扩散模型相比基于像素的模型至少带来2.7×倍的加速，同时将FID分数提升了至少1.6×倍。

与表7中其他修复方法的比较显示，我们引入注意力机制的模型在FID指标上优于文献[88]，整体图像质量有所提升。未遮挡图像与我们生成样本之间的LPIPS值略高于文献[88]。我们认为这是因为文献[88]仅生成单一结果，倾向于恢复出更接近平均图像的输出，而我们的LDM {v*} 能产生多样化结果（图21）。此外，在用户研究（表4）中，受试者对我们生成结果的偏好程度也高于文献[88]。

基于这些初步结果，我们还在无注意力机制的*VQ*正则化第一阶段潜空间中训练了一个更大的扩散模型（即表7中的*big*）。参照[15]的方法，该扩散模型的UNet在其特征层次结构的三个层级上使用注意力层，采用BigGAN[3]残差块进行上采样和下采样，并包含3.87亿参数。

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.

图11. 使用我们的*big, w/ ft*修复模型进行物体移除的定性结果。更多结果见图22。

而非215M。训练后，我们注意到在256^2和512^2分辨率下生成的样本质量存在差异，我们推测这是由额外的注意力模块引起的。然而，在512^2分辨率下对模型进行半个周期的微调，使其能够适应新的特征统计量，并在图像修复任务上创造了新的FID最佳纪录（见表7中的*big, w/o attn, w/ ft*，图11）。

5. 局限性与社会影响

局限性　尽管与基于像素的方法相比LDMs显著降低了计算需求，但其顺序采样过程仍比GANs慢。此外，当需要高精度时，使用LDMs可能存在问题：虽然在我们{v*}4自编码模型中图像质量损失非常小（见图1），但其重建能力可能成为需要在像素空间中实现细粒度精度的任务的瓶颈。我们认为，我们的超分辨率模型（第4.4节）在这方面已经存在一定限制。

社会影响　生成式媒体模型（如图像生成）是一把双刃剑：一方面，它们

Method	40-50% masked FID ↓	40-50% masked LPIPS ↓	All samples FID ↓	All samples 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

表7. 在Places [108]测试图像中30k个尺寸为512 × 512的裁剪区域上进行修复性能比较。*40- 50%*列报告了在困难样本（需修复图像区域40-50%）上计算的指标。由于[88]中使用的原始测试集不可用，†在我们的测试集上重新计算。

实现各种创意应用，特别是像我们这样降低训练和推理成本的方法，有可能促进该技术的普及并使其探索民主化。另一方面，这也意味着创建和传播被篡改的数据或散布虚假信息和垃圾内容变得更加容易。特别是，对图像的蓄意操纵（“深度伪造”）在此背景下是一个普遍问题，而女性尤其受到其不成比例的影响[13, 24]。

生成模型也可能泄露其训练数据[5, 90]，当数据包含敏感或个人隐私信息且未经明确同意收集时，这一问题尤为令人担忧。然而，这种情况在多大程度上也适用于图像{v*}的扩散模型，目前尚未完全明确。

最后，深度学习模块倾向于复制或加剧数据中已经存在的偏见[22, 38, 91]。虽然扩散模型比{v*}基于GAN的方法能更好地覆盖数据分布，但我们的两阶段方法结合了对抗训练和基于似然的目标，其在多大程度上歪曲了数据仍然是一个重要的研究问题。

关于深度生成模型伦理考量的更全面、详细讨论，请参阅*e.g.*。[13]。

6. 结论

我们提出了潜在扩散模型，这是一种简单而高效的方法，能在不降低质量的前提下显著提升去噪扩散模型的训练和采样效率。基于此以及我们的交叉注意力条件机制，实验表明，在广泛的图像条件合成任务中，我们的方法无需特定任务架构即可取得优于当前先进技术的结果。

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.

References

[1] Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1122–1131. IEEE Computer Society, 2017. 1

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan, 2017. 3

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *Int. Conf. Learn. Represent.*, 2019. 1, 2, 7, 8, 22, 28

[4] Holger Caesar, Jasper R. R. Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1209–1218. Computer Vision Foundation / IEEE Computer Society, 2018. 7, 20, 22

[5] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650, 2021. 9

[6] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 1691–1703. PMLR, 2020. 3

[7] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J. Weiss, Mohammad Norouzi, and William Chan. Wavegrad: Estimating gradients for waveform generation. In *ICLR*. OpenReview.net, 2021. 1

[8] Lu Chi, Borui Jiang, and Yadong Mu. Fast fourier convolution. In *NeurIPS*, 2020. 8

[9] Rewon Child. Very deep vaes generalize autoregressive models and can outperform them on images. *CoRR*, abs/2011.10650, 2020. 3

[10] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *CoRR*, abs/1904.10509, 2019. 3

[11] Bin Dai and David P. Wipf. Diagnosing and enhancing VAE models. In *ICLR (Poster)*. OpenReview.net, 2019. 2, 3

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE Computer Society, 2009. 1, 5, 7, 22

[13] Emily Denton. Ethical considerations of generative ai. AI for Content Creation Workshop, CVPR, 2021. 9

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. 7

[15] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *CoRR*, abs/2105.05233, 2021. 1, 2, 3, 4, 6, 7, 8, 18, 22, 25, 26, 28

[16] Sander Dieleman. Musings on typicality, 2020. 1, 3

[17] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. Cogview: Mastering text-to-image generation via transformers. *CoRR*, abs/2105.13290, 2021. 6, 7

[18] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation, 2015. 3

[19] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. 1, 3

[20] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Adv. Neural Inform. Process. Syst.*, pages 658–666, 2016. 3

[21] Patrick Esser, Robin Rombach, Andreas Blattmann, and Björn Ommer. Imagebart: Bidirectional context with multinomial diffusion for autoregressive image synthesis. *CoRR*, abs/2108.08827, 2021. 6, 7, 22

[22] Patrick Esser, Robin Rombach, and Björn Ommer. A note on data biases in generative models. *arXiv preprint arXiv:2012.02516*, 2020. 9

[23] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. *CoRR*, abs/2012.09841, 2020. 2, 3, 4, 6, 7, 21, 22, 29, 34, 36

[24] Mary Anne Franks and Ari Ezra Waldman. Sex, lies, and videotape: Deep fakes and free speech delusions. *Md. L. Rev.*, 78:892, 2018. 9

[25] Kevin Frans, Lisa B. Soros, and Olaf Witkowski. Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. *ArXiv*, abs/2106.14843, 2021. 3

[26] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. *CoRR*, abs/2203.13131, 2022. 6, 7, 16

[27] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial networks. *CoRR*, 2014. 1, 2

[28] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans, 2017. 3

[29] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Adv. Neural Inform. Process. Syst.*, pages 6626–6637, 2017. 1, 5, 26

[30] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 1, 2, 3, 4, 6, 17

[31] Jonathan Ho, Chitwan Saharia, William Chan, David J. Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *CoRR*, abs/2106.15282, 2021. 1, 3, 22

参考文献

[1] Eirikur Agustsson 与 Radu Timofte。NTIRE 2017 单图像超分辨率挑战赛：数据集与研究。收录于 *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2017, Honolulu, HI, USA, July 21-26, 2017*，第 1122–1131 页。IEEE 计算机学会，2017 年。1[2] Martin Arjovsky, Soumith Chintala 与 Léon Bottou。Wasserstein GAN，2017 年。3[3] Andrew Brock, Jeff Donahue 与 Karen Simonyan。面向高保真自然图像合成的大规模 GAN 训练。收录于 *Int. Conf. Learn. Represent.*，2019 年。1, 2, 7, 8, 22, 28[4] Holger Caesar, Jasper R. R. Uijlings 与 Vittorio Ferrari。Coco-stuff：上下文中的物体与场景类别。收录于 *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*，第 1209–1218 页。计算机视觉基金会 / IEEE 计算机学会，2018 年。7, 20, 22[5] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson 等。从大型语言模型中提取训练数据。收录于 *30th USENIX Security Symposium (USENIX Security 21)*，第 2633–2650 页，2021 年。9[6] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Hee- woo Jun, David Luan 与 Ilya Sutskever。基于像素的生成式预训练。收录于 *ICML*，第 119 卷 *Proceedings of Machine Learning Research*，第 1691–1703 页。PMLR，2020 年。3[7] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J. Weiss, Mo- hammad Norouzi 与 William Chan。Wavegrad：用于波形生成的梯度估计。收录于 *ICLR*。OpenReview.net，2021 年。1[8] Lu Chi, Borui Jiang 与 Yadong Mu。快速傅里叶卷积。收录于 *NeurIPS*，2020 年。8[9] Rewon Child。极深 VAE 泛化自回归模型并能在图像上超越它们。*CoRR*，abs/2011.10650，2020 年。3[10] Rewon Child, Scott Gray, Alec Radford 与 Ilya Sutskever。使用稀疏 Transformer 生成长序列。*CoRR*，abs/1904.10509，2019 年。3[11] Bin Dai 与 David P. Wipf。诊断与增强 VAE 模型。收录于 *ICLR (Poster)*。OpenReview.net，2019 年。2, 3[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li 与 Fei-Fei Li。Imagenet：一个大规模分层图像数据库。收录于 *CVPR*，第 248–255 页。IEEE 计算机学会，2009 年。1, 5, 7, 22[13] Emily Denton。生成式 AI 的伦理考量。AI 内容创作研讨会，CVPR，2021 年。9[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee 与 Kristina Toutanova。BERT：用于语言理解的深度双向 Transformer 预训练。*CoRR*，abs/1810.04805，2018 年。7[15] Prafulla Dhariwal 与 Alex Nichol。扩散模型在图像合成上击败 GAN。*CoRR*，abs/2105.05233，2021 年。1, 2, 3, 4, 6, 7, 8, 18, 22, 25, 26, 28

[16] Sander Dieleman。关于典型性的思考，2020年。1, 3[17] 丁铭、杨卓毅、洪文义、郑文迪、周昶、尹达、林俊旸、邹旭、邵舟、杨红霞、唐杰。Cogview：通过Transformer掌握文本到图像生成。*CoRR*，abs/2105.13290，2021年。6, 7[18] Laurent Dinh、David Krueger、Yoshua Bengio。NICE：非线性独立分量估计，2015年。3[19] Laurent Dinh、Jascha Sohl-Dickstein、Samy Bengio。使用Real NVP进行密度估计。收录于*5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*。OpenReview.net，2017年。1, 3[20] Alexey Dosovitskiy、Thomas Brox。使用基于深度网络的感知相似性度量生成图像。收录于Daniel D. Lee、Masashi Sugiyama、Ulrike von Luxburg、Isabelle Guyon、Roman Garnett编辑的 *Adv. Neural Inform. Process. Syst.*，第658–666页，2016年。3[21] Patrick Esser、Robin Rombach、Andreas Blattmann、Björn Ommer。ImageBART：用于自回归图像合成的具有多项扩散的双向上下文。*CoRR*，abs/2108.08827，2021年。6, 7, 22[22] Patrick Esser、Robin Rombach、Björn Ommer。关于生成模型中数据偏差的说明。*arXiv preprint arXiv:2012.02516*，2020年。9[23] Patrick Esser、Robin Rombach、Björn Ommer。驯服Transformer用于高分辨率图像合成。*CoRR*，abs/2012.09841，2020年。2, 3, 4, 6, 7, 21, 22, 29, 34, 36[24] Mary Anne Franks、Ari Ezra Waldman。性、谎言和录像带：深度伪造与言论自由的错觉。*Md. L. Rev.*，78:892，2018年。9[25] Kevin Frans、Lisa B. Soros、Olaf Witkowski。ClipDraw：通过语言-图像编码器探索文本到绘图合成。*ArXiv*，abs/2106.14843，2021年。3[26] Oran Gafni、Adam Polyak、Oron Ashual、Shelly Sheynin、Devi Parikh、Yaniv Taigman。Make-A-Scene：基于场景和人类先验的文本到图像生成。*CoRR*，abs/2203.13131，2022年。6, 7, 16[27] Ian J. Goodfellow、Jean Pouget-Abadie、Mehdi Mirza、Bing Xu、David Warde-Farley、Sherjil Ozair、Aaron C. Courville、Yoshua Bengio。生成对抗网络。*CoRR*，2014年。1, 2[28] Ishaan Gulrajani、Faruk Ahmed、Martin Arjovsky、Vincent Dumoulin、Aaron Courville。Wasserstein GANs的改进训练，2017年。3[29] Martin Heusel、Hubert Ramsauer、Thomas Unterthiner、Bernhard Nessler、Sepp Hochreiter。通过双时间尺度更新规则训练的GAN收敛到局部纳什均衡。收录于 *Adv. Neural Inform. Process. Syst.*，第6626–6637页，2017年。1, 5, 26[30] Jonathan Ho、Ajay Jain、Pieter Abbeel。去噪扩散概率模型。收录于*NeurIPS*，2020年。1, 2, 3, 4, 6, 17[31] Jonathan Ho、Chitwan Saharia、William Chan、David J. Fleet、Mohammad Norouzi、Tim Salimans。用于高保真图像生成的级联扩散模型。*CoRR*，abs/2106.15282，2021年。1, 3, 22

[32] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 6, 7, 16, 22, 28, 37, 38

[33] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 5967–5976. IEEE Computer Society, 2017. 3, 4

[34] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976, 2017. 4

[35] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier J. Hénaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver IO: A general architecture for structured inputs &outputs. *CoRR*, abs/2107.14795, 2021. 4

[36] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and João Carreira. Perceiver: General perception with iterative attention. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 4651–4664. PMLR, 2021. 4, 5

[37] Manuel Jahn, Robin Rombach, and Björn Ommer. High-resolution complex scene synthesis with transformers. *CoRR*, abs/2105.06458, 2021. 20, 22, 27

[38] Niharika Jain, Alberto Olmo, Sailik Sengupta, Lydia Manikonda, and Subbarao Kambhampati. Imperfect imaganation: Implications of gans exacerbating biases on facial data augmentation and snapchat selfie lenses. *arXiv preprint arXiv:2001.09528*, 2020. 9

[39] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *CoRR*, abs/1710.10196, 2017. 5, 6

[40] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4401–4410, 2019. 1

[41] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5, 6

[42] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. *CoRR*, abs/1912.04958, 2019. 2, 6, 28

[43] Dongjun Kim, Seungjae Shin, Kyungwoo Song, Wanmo Kang, and Il-Chul Moon. Score matching model for unbounded data score. *CoRR*, abs/2106.05527, 2021. 6

[44] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, 2018. 3

[45] Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *CoRR*, abs/2107.00630, 2021. 1, 3, 16

[46] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR*, 2014. 1, 3, 4, 29

[47] Zhifeng Kong and Wei Ping. On fast sampling of diffusion probabilistic models. *CoRR*, abs/2106.00132, 2021. 3

[48] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *ICLR*. OpenReview.net, 2021. 1

[49] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, and Vittorio Ferrari. The open images dataset V4: unified image classification, object detection, and visual relationship detection at scale. *CoRR*, abs/1811.00982, 2018. 7, 20, 22

[50] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *CoRR*, abs/1904.06991, 2019. 5, 26

[51] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. 6, 7, 27

[52] Yuqing Ma, Xianglong Liu, Shihao Bai, Le-Yi Wang, Aishan Liu, Dacheng Tao, and Edwin Hancock. Region-wise generative adversarial imageinpainting for large missing areas. *ArXiv*, abs/1909.12507, 2019. 9

[53] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *CoRR*, abs/2108.01073, 2021. 1

[54] Lars M. Mescheder. On the convergence properties of GAN training. *CoRR*, abs/1801.04406, 2018. 3

[55] Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. Unrolled generative adversarial networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. 3

[56] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014. 4

[57] Gautam Mittal, Jesse H. Engel, Curtis Hawthorne, and Ian Simon. Symbolic music generation with diffusion models. *CoRR*, abs/2103.16091, 2021. 1

[58] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Z. Qureshi, and Mehran Ebrahimi. Edgeconnect: Generative image inpainting with adversarial edge learning. *ArXiv*, abs/1901.00212, 2019. 9

[59] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. *CoRR*, abs/2112.10741, 2021. 6, 7, 16

[60] Anton Obukhov, Maximilian Seitzer, Po-Wei Wu, Semen Zhydenko, Jonathan Kyl, and Elvis Yu-Jing Lin.

[32] Jonathan Ho 与 Tim Salimans。无分类器扩散引导。于 *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*，2021年。6, 7, 16, 22, 28, 37, 38[33] Phillip Isola、Jun-Yan Zhu、Tinghui Zhou 与 Alexei A. Efros。基于条件对抗网络的图像到图像翻译。于 *CVPR*，第5967–5976页。IEEE计算机学会，2017年。3, 4[34] Phillip Isola、Jun-Yan Zhu、Tinghui Zhou 与 Alexei A. Efros。基于条件对抗网络的图像到图像翻译。*2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*，第5967–5976页，2017年。4[35] Andrew Jaegle、Sebastian Borgeaud、Jean-Baptiste Alayrac、Carl Doersch、Catalin Ionescu、David Ding、Skanda Koppula、Daniel Zoran、Andrew Brock、Evan Shelhamer、Olivier J. Hénaff、Matthew M. Botvinick、Andrew Zisserman、Oriol Vinyals 与 João Carreira。Perceiver IO：一种用于结构化输入和输出的通用架构。*CoRR*，abs/2107.14795，2021年。4[36] Andrew Jaegle、Felix Gimeno、Andy Brock、Oriol Vinyals、Andrew Zisserman 与 João Carreira。Perceiver：通过迭代注意力实现通用感知。载于 Marina Meila 与 Tong Zhang 编辑的 *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*，第139卷 *Proceedings of Machine Learning Research*，第4651–4664页。PMLR，2021年。4, 5[37] Manuel Jahn、Robin Rombach 与 Björn Ommer。基于Transformer的高分辨率复杂场景合成。*CoRR*，abs/2105.06458，2021年。20, 22, 27[38] Niharika Jain、Alberto Olmo、Sailik Sengupta、Lydia Manikonda 与 Subbarao Kambhampati。不完美的想象：GANs加剧面部数据增强和Snapchat自拍镜头偏见的启示。*arXiv preprint arXiv:2001.09528*，2020年。9[39] Tero Karras、Timo Aila、Samuli Laine 与 Jaakko Lehtinen。渐进式增长GANs以提升质量、稳定性和多样性。*CoRR*，abs/1710.10196，2017年。5, 6[40] Tero Karras、Samuli Laine 与 Timo Aila。一种用于生成对抗网络的基于风格的生成器架构。于 *IEEE Conf. Comput. Vis. Pattern Recog.*，第4401–4410页，2019年。1[41] T. Karras、S. Laine 与 T. Aila。一种用于生成对抗网络的基于风格的生成器架构。于 *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*，2019年。5, 6[42] Tero Karras、Samuli Laine、Miika Aittala、Janne Hellsten、Jaakko Lehtinen 与 Timo Aila。分析与改进StyleGAN的图像质量。*CoRR*，abs/1912.04958，2019年。2, 6, 28[43] Dongjun Kim、Seungjae Shin、Kyungwoo Song、Wanmo Kang 与 Il-Chul Moon。无界数据分数的得分匹配模型。*CoRR*，abs/2106.05527，2021年。6[44] Durk P Kingma 与 Prafulla Dhariwal。Glow：具有可逆1x1卷积的生成流。载于 S. Bengio、H. Wallach、H. Larochelle、K. Grauman、N. Cesa-Bianchi 与 R. Garnett 编辑的 *Advances in Neural Information Process- ing Systems*，2018年。3[45] Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. 变分扩散模型。*CoRR*, abs/2107.00630, 2021. 1, 3, 16[46] Diederik P. Kingma and Max Welling. 自动编码变分贝叶斯。于 *2nd International Conference on Learning Representations, ICLR*, 2014. 1, 3, 4, 29[47] Zhifeng Kong and Wei Ping. 论扩散概率模型的快速采样。*CoRR*, abs/2106.00132, 2021. 3[48] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave：一种用于音频合成的通用扩散模型。于 *ICLR*. OpenReview.net, 2021. 1[49] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, and Vittorio Ferrari. Open Images 数据集 V4：大规模统一图像分类、目标检测与视觉关系检测。*CoRR*, abs/1811.00982, 2018. 7, 20, 22[50] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. 用于评估生成模型的改进精确率与召回率指标。*CoRR*, abs/1904.06991, 2019. 5, 26[51] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO：上下文中的常见物体。*CoRR*, abs/1405.0312, 2014. 6, 7, 27[52] Yuqing Ma, Xianglong Liu, Shihao Bai, Le-Yi Wang, Aishan Liu, Dacheng Tao, and Edwin Hancock. 面向大面积缺失的区域生成对抗图像修复。*ArXiv*, abs/1909.12507, 2019. 9[53] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit：基于随机微分方程的图像合成与编辑。*CoRR*, abs/2108.01073, 2021. 1[54] Lars M. Mescheder. 论 GAN 训练的收敛性。*CoRR*, abs/1801.04406, 2018. 3[55] Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. 展开生成对抗网络。于 *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. 3[56] Mehdi Mirza and Simon Osindero. 条件生成对抗网络。*CoRR*, abs/1411.1784, 2014. 4[57] Gautam Mittal, Jesse H. Engel, Curtis Hawthorne, and Ian Simon. 基于扩散模型的符号音乐生成。*CoRR*, abs/2103.16091, 2021. 1[58] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Z. Qureshi, and Mehran Ebrahimi. Edgeconnect：基于对抗边缘学习的生成图像修复。*ArXiv*, abs/1901.00212, 2019. 9[59] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE：迈向基于文本引导扩散模型的逼真图像生成与编辑。*CoRR*, abs/2112.10741, 2021. 6, 7, 16[60] Anton Obukhov, Maximilian Seitzer, Po-Wei Wu, Semen Zhydenko, Jonathan Kyl, and Elvis Yu-Jing Lin.

High-fidelity performance metrics for generative models in pytorch, 2020. Version: 0.3.0, DOI: 10.5281/zenodo.4957738. 26, 27

[61] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 4, 7

[62] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 22

[63] Gaurav Parmar, Dacheng Li, Kwonjoon Lee, and Zhuowen Tu. Dual contradistinctive generative autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 823–832. Computer Vision Foundation / IEEE, 2021. 6

[64] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On buggy resizing libraries and surprising subtleties in fid calculation. *arXiv preprint arXiv:2104.11222*, 2021. 26

[65] David A. Patterson, Joseph Gonzalez, Quoc V. Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David R. So, Maud Texier, and Jeff Dean. Carbon emissions and large neural network training. *CoRR*, abs/2104.10350, 2021. 2

[66] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *CoRR*, abs/2102.12092, 2021. 1, 2, 3, 4, 7, 21, 27

[67] Ali Razavi, Aäron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with VQ-VAE-2. In *NeurIPS*, pages 14837–14847, 2019. 1, 2, 3, 22

[68] Scott E. Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *ICML*, 2016. 4

[69] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on International Conference on Machine Learning, ICML*, 2014. 1, 4, 29

[70] Robin Rombach, Patrick Esser, and Björn Ommer. Network-to-network translation with conditional invertible neural networks. In *NeurIPS*, 2020. 3

[71] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI (3)*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241. Springer, 2015. 2, 3, 4

[72] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *CoRR*, abs/2104.07636, 2021. 1, 4, 8, 16, 22, 23, 27

[73] Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P. Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *CoRR*, abs/1701.05517, 2017. 1, 3

[74] Dave Salvator. NVIDIA Developer Blog. https://developer.nvidia.com/blog/getting-immediate-speedups-with-a100-tf32, 2020. 28

[75] Robin San-Roman, Eliya Nachmani, and Lior Wolf. Noise estimation for generative diffusion models. *CoRR*, abs/2104.02600, 2021. 3

[76] Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected gans converge faster. *CoRR*, abs/2111.01007, 2021. 6

[77] Edgar Schönfeld, Bernt Schiele, and Anna Khoreva. A u-net based discriminator for generative adversarial networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8204–8213. Computer Vision Foundation / IEEE, 2020. 6

[78] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs, 2021. 6, 7

[79] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *Int. Conf. Learn. Represent.*, 2015. 29, 43, 44, 45

[80] Abhishek Sinha, Jiaming Song, Chenlin Meng, and Stefano Ermon. D2C: diffusion-denoising models for few-shot conditional generation. *CoRR*, abs/2106.06819, 2021. 3

[81] Charlie Snell. Alien Dreams: An Emerging Art Scene. https://ml.berkeley.edu/blog/posts/clip-art/, 2021. [Online; accessed November-2021]. 2

[82] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. *CoRR*, abs/1503.03585, 2015. 1, 3, 4, 18

[83] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. 4

[84] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*. OpenReview.net, 2021. 3, 5, 6, 22

[85] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *CoRR*, abs/2011.13456, 2020. 1, 3, 4, 18

[86] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for modern deep learning research. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 13693–13696. AAAI Press, 2020. 2

PyTorch中生成模型的高保真性能指标，2020年。版本：0.3.0，DOI：10.5281/zen-odo.4957738。26, 27 [61] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. 基于空间自适应归一化的语义图像合成。发表于 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 4, 7 [62] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. 基于空间自适应归一化的语义图像合成。发表于 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019年6月. 22 [63] Gaurav Parmar, Dacheng Li, Kwonjoon Lee, and Zhuowen Tu. 双重对比生成自编码器。发表于 *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 第823–832页. 计算机视觉基金会 / IEEE, 2021. 6 [64] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. 关于有缺陷的图像缩放库与FID计算中令人惊讶的微妙之处。*arXiv preprint arXiv:2104.11222*, 2021. 26 [65] David A. Patterson, Joseph Gonzalez, Quoc V. Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David R. So, Maud Texier, and Jeff Dean. 碳排放与大型神经网络训练。*CoRR*, abs/2104.10350, 2021. 2 [66] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 零样本文本到图像生成。*CoRR*, abs/2102.12092, 2021. 1, 2, 3, 4, 7, 21, 27 [67] Ali Razavi, Aäron van den Oord, and Oriol Vinyals. 使用VQ-VAE-2生成多样化的高保真图像。发表于 *NeurIPS*, 第14837–14847页, 2019. 1, 2, 3, 22 [68] Scott E. Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 生成对抗式文本到图像合成。发表于 *ICML*, 2016. 4 [69] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 深度生成模型中的随机反向传播与近似推断。发表于 *Proceedings of the 31st International Conference on International Conference on Machine Learning, ICML*, 2014. 1, 4, 29 [70] Robin Rombach, Patrick Esser, and Björn Ommer. 使用条件可逆神经网络的网络到网络翻译。发表于 *NeurIPS*, 2020. 3 [71] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net：用于生物医学图像分割的卷积网络。发表于 *MICCAI (3)*, 卷9351, *Lecture Notes in Computer Science*, 第234–241页. Springer, 2015. 2, 3, 4 [72] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. 通过迭代细化的图像超分辨率。*CoRR*, abs/2104.07636, 2021. 1, 4, 8, 16, 22, 23, 27 [73] Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P. Kingma. PixelCNN++：使用离散逻辑混合似然及其他改进的PixelCNN。*CoRR*, abs/1701.05517, 2017. 1, 3 [74] Dave Salvator. NVIDIA开发者博客。https : / / developer . nvidia . com / blog / getting - 即时加速-A100-TF32，2020年。28 [75] Robin San-Roman、Eliya Nachmani和Lior Wolf。生成扩散模型的噪声估计。*CoRR*，abs/2104.02600，2021年。3 [76] Axel Sauer、Kashyap Chitta、Jens Müller和Andreas Geiger。投影GAN收敛更快。*CoRR*，abs/2111.01007，2021年。6 [77] Edgar Schönfeld、Bernt Schiele和Anna Khoreva。基于U-Net的生成对抗网络判别器。收录于 *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*，第8204–8213页。计算机视觉基金会/IEEE，2020年。6 [78] Christoph Schuhmann、Richard Vencu、Romain Beaumont、Robert Kaczmarczyk、Clayton Mullis、Aarush Katta、Theo Coombes、Jenia Jitsev和Aran Komatsuzaki。LAION-400M：包含4亿个图文对的CLIP过滤开源数据集，2021年。6、7 [79] Karen Simonyan和Andrew Zisserman。用于大规模图像识别的极深度卷积网络。收录于Yoshua Bengio和Yann LeCun编辑的*Int. Conf. Learn. Represent.*，2015年。29、43、44、45 [80] Abhishek Sinha、Jiaming Song、Chenlin Meng和Stefano Ermon。D2C：用于少样本条件生成的扩散去噪模型。*CoRR*，abs/2106.06819，2021年。3 [81] Charlie Snell。异梦：新兴艺术场景。https://ml.berkeley.edu/blog/posts/clip-art/，2021年。[在线访问于2021年11月]。2 [82] Jascha Soh l-Dickstein、Eric A. Weiss、Niru Maheswaranathan和Surya Ganguli。基于非平衡热力学的深度无监督学习。*CoRR*，abs/1503.03585，2015年。1、3、4、18 [83] Kihyuk Sohn、Honglak Lee和Xinchen Yan。使用深度条件生成模型学习结构化输出表示。收录于C. Cortes、N. Lawrence、D. Lee、M. Sugiyama和R. Garnett编辑的*Advances in Neural Information Processing Systems*，第28卷。Curran Associates, Inc.，2015年。4 [84] Jiaming Song、Chenlin Meng和Stefano Ermon。去噪扩散隐式模型。收录于*ICLR*。OpenReview.net，2021年。3、5、6、22 [85] Yang Song、Jascha Sohl-Dickstein、Diederik P. Kingma、Abhishek Kumar、Stefano Ermon和Ben Poole。通过随机微分方程进行基于分数的生成建模。*CoRR*，abs/2011.13456，2020年。1、3、4、18 [86] Emma Strubell、Ananya Ganesh和Andrew McCallum。现代深度学习研究的能源与政策考量。收录于 *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*，第13693–13696页。AAAI Press，2020年。2

[87] Wei Sun and Tianfu Wu. Learning layout and style reconfigurable gans for controllable image synthesis. *CoRR*, abs/2003.11571, 2020. 22, 27

[88] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor S. Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. *ArXiv*, abs/2109.07161, 2021. 8, 9, 26, 32

[89] Tristan Sylvain, Pengchuan Zhang, Yoshua Bengio, R. Devon Hjelm, and Shikhar Sharma. Object-centric image generation from layouts. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 2647–2655. AAAI Press, 2021. 20, 22, 27

[90] Patrick Tinsley, Adam Czajka, and Patrick Flynn. This face does not exist... but it might be yours! identity leakage in generative models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1320–1328, 2021. 9

[91] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011. 9

[92] Arash Vahdat and Jan Kautz. NVAE: A deep hierarchical variational autoencoder. In *NeurIPS*, 2020. 3

[93] Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. *CoRR*, abs/2106.05931, 2021. 2, 3, 5, 6

[94] Aaron van den Oord, Nal Kalchbrenner, Lasse Espeholt, koray kavukcuoglu, Oriol Vinyals, and Alex Graves. Conditional image generation with pixelcnn decoders. In *Advances in Neural Information Processing Systems*, 2016. 3

[95] Aäron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. *CoRR*, abs/1601.06759, 2016. 3

[96] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *NIPS*, pages 6306–6315, 2017. 2, 4, 29

[97] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017. 3, 4, 5, 7

[98] Rivers Have Wings. Tweet on Classifier-free guidance for autoregressive models. https://twitter.com/RiversHaveWings/status/1478093658716966912, 2022. 6

[99] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771, 2019. 26

[100] Zhisheng Xiao, Karsten Kreis, Jan Kautz, and Arash Vahdat. VAEBM: A symbiosis between variational autoencoders and energy-based models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 6

[101] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using VQ-VAE and transformers. *CoRR*, abs/2104.10157, 2021. 3

[102] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. LSUN: construction of a large-scale image dataset using deep learning with humans in the loop. *CoRR*, abs/1506.03365, 2015. 5

[103] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan, 2021. 3, 4

[104] Jiahui Yu, Zhe L. Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Free-form image inpainting with gated convolution. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4470–4479, 2019. 9

[105] K. Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. *ArXiv*, abs/2103.14006, 2021. 23

[106] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 3, 8, 19

[107] Shengyu Zhao, Jianwei Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I-Chao Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks. *ArXiv*, abs/2103.10428, 2021. 9

[108] Bolei Zhou, Àgata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:1452–1464, 2018. 8, 9, 26

[109] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. LAFITE: towards language-free training for text-to-image generation. *CoRR*, abs/2111.13792, 2021. 6, 7, 16

[87] 孙伟与吴天福。学习布局与风格可重构的生成对抗网络以实现可控图像合成。*CoRR*，abs/2003.11571，2020年。22, 27[88] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, 与 Victor S. Lempitsky。使用傅里叶卷积进行分辨率鲁棒的大掩码修复。*ArXiv*，abs/2109.07161，2021年。8, 9, 26, 32[89] Tristan Sylvain, Pengchuan Zhang, Yoshua Bengio, R. Devon Hjelm, 与 Shikhar Sharma。从布局生成以对象为中心的图像。收录于 *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*，第2647–2655页。AAAI出版社，2021年。20, 22, 27[90] Patrick Tinsley, Adam Czajka, 与 Patrick Flynn。这张脸不存在……但它可能是你的！生成模型中的身份泄露。收录于 *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*，第1320–1328页，2021年。9[91] Antonio Torralba 与 Alexei A Efros。对数据集偏差的无偏见审视。收录于 *CVPR 2011*，第1521–1528页。IEEE，2011年。9[92] Arash Vahdat 与 Jan Kautz。NVAE：一种深度分层变分自编码器。收录于 *NeurIPS*，2020年。3[93] Arash Vahdat, Karsten Kreis, 与 Jan Kautz。潜在空间中的基于分数的生成建模。*CoRR*，abs/2106.05931，2021年。2, 3, 5, 6[94] Aaron van den Oord, Nal Kalchbrenner, Lasse Espeholt, koray kavukcuoglu, Oriol Vinyals, 与 Alex Graves。使用PixelCNN解码器进行条件图像生成。收录于 *Advances in Neural Information Processing Systems*，2016年。3[95] Aäron van den Oord, Nal Kalchbrenner, 与 Koray Kavukcuoglu。像素循环神经网络。*CoRR*，abs/1601.06759，2016年。3[96] Aäron van den Oord, Oriol Vinyals, 与 Koray Kavukcuoglu。神经离散表示学习。收录于 *NIPS*，第6306–6315页，2017年。2, 4, 29[97] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, 与 Illia Polosukhin。注意力机制就是你所需要的一切。收录于 *NIPS*，第5998–6008页，2017年。3, 4, 5, 7[98] Rivers Have Wings。关于自回归模型的无分类器引导的推文。https : / / twitter . com / RiversHaveWings / status / 1478093658716966912，2022年。6[99] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, 与 Jamie Brew。Huggingface的Transformers：最先进的自然语言处理。*CoRR*，abs/1910.03771，2019年。26[100] Zhisheng Xiao, Karsten Kreis, Jan Kautz, 与 Arash Vahdat。VAEBM：变分自编码器与基于能量的模型之间的共生关系。收录于 *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*[101] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, 与 Aravind Srinivas。Videogpt: 使用 VQ-VAE 和 Transformer 生成视频。*CoRR*, abs/2104.10157, 2021. 3[102] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, 与 Jianxiong Xiao。LSUN: 利用人在回路的深度学习构建大规模图像数据集。*CoRR*, abs/1506.03365, 2015. 5[103] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, 与 Yonghui Wu。使用改进的 VQGAN 进行矢量量化图像建模，2021. 3, 4[104] Jiahui Yu, Zhe L. Lin, Jimei Yang, Xiaohui Shen, Xin Lu, 与 Thomas S. Huang。使用门控卷积进行自由形式图像修复。*2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 页码 4470–4479, 2019. 9[105] K. Zhang, Jingyun Liang, Luc Van Gool, 与 Radu Timofte。为深度盲图像超分辨率设计实用的退化模型。*ArXiv*, abs/2103.14006, 2021. 23[106] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, 与 Oliver Wang。深度特征作为感知度量标准的惊人有效性。收录于 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018年6月. 3, 8, 19[107] Shengyu Zhao, Jianwei Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I-Chao Chang, 与 Yan Xu。通过协同调制生成对抗网络进行大规模图像补全。*ArXiv*, abs/2103.10428, 2021. 9[108] Bolei Zhou, Àgata Lapedriza, Aditya Khosla, Aude Oliva, 与 Antonio Torralba。Places: 一个用于场景识别的千万级图像数据库。*IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:1452–1464, 2018. 8, 9, 26[109] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, 与 Tong Sun。LAFITE: 迈向无需语言的文本到图像生成训练。*CoRR*, abs/2111.13792, 2021. 6, 7, 16

Appendix

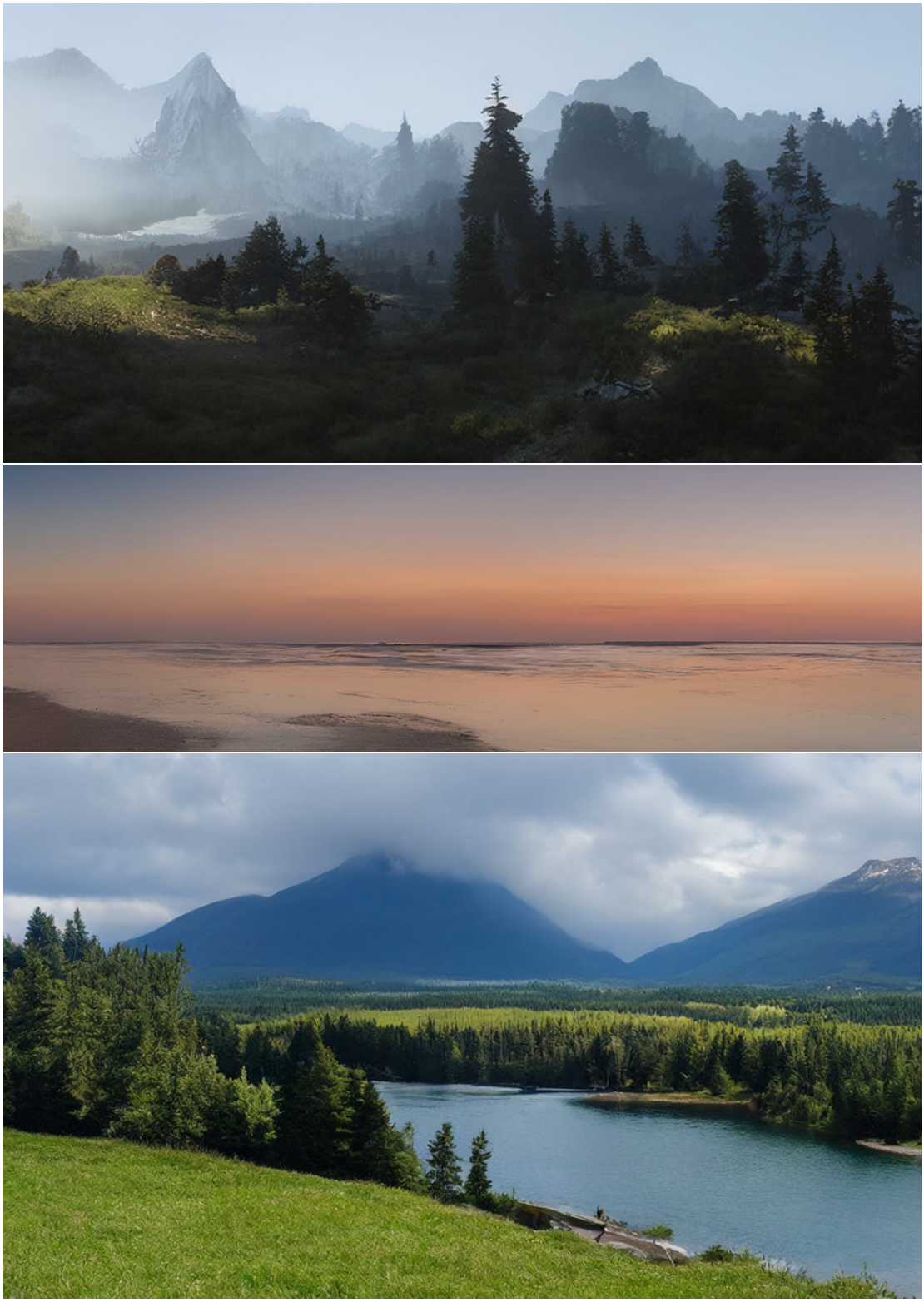

Figure 12. Convolutional samples from the semantic landscapes model as in Sec. 4.3.2, finetuned on 512^2 images.

附录

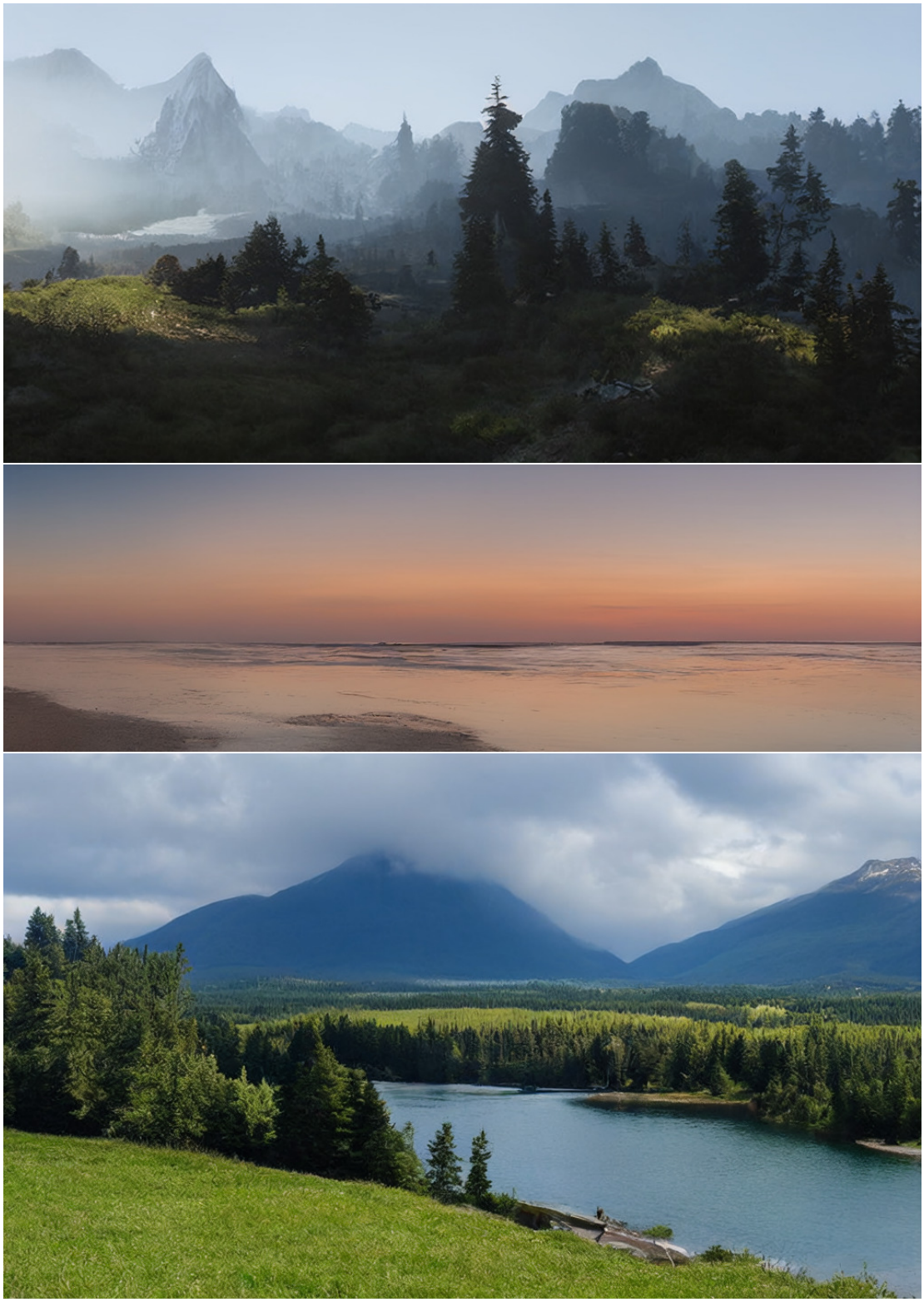

图12. 如第4.3.2节所述，在512^2图像上微调的语义景观模型的卷积样本。

'A painting of the last supper by Picasso.'

'An oil painting of a latent space.'

'An epic painting of Gandalf the Black summoning thunder and lightning in the mountains.'

'A sunset over a mountain range, vector image.'

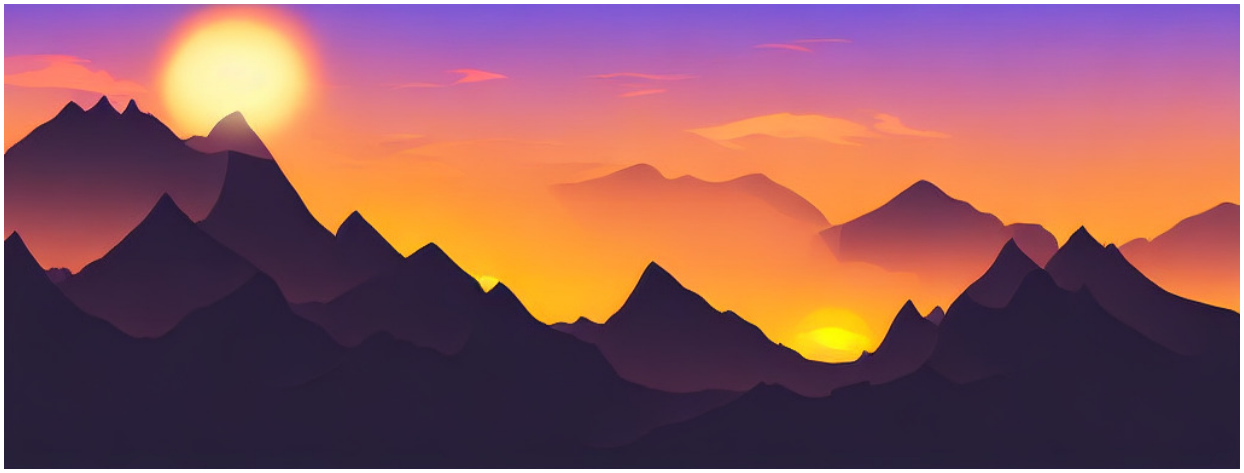

Figure 13. Combining classifier free diffusion guidance with the convolutional sampling strategy from Sec. 4.3.2, our 1.45B parameter text-to-image model can be used for rendering images larger than the native 256^2 resolution the model was trained on.

'A painting of the last supper by Picasso.'

'An oil painting of a latent space.'

'An epic painting of Gandalf the Black summoning thunder and lightning in the mountains.'

'A sunset over a mountain range, vector image.'

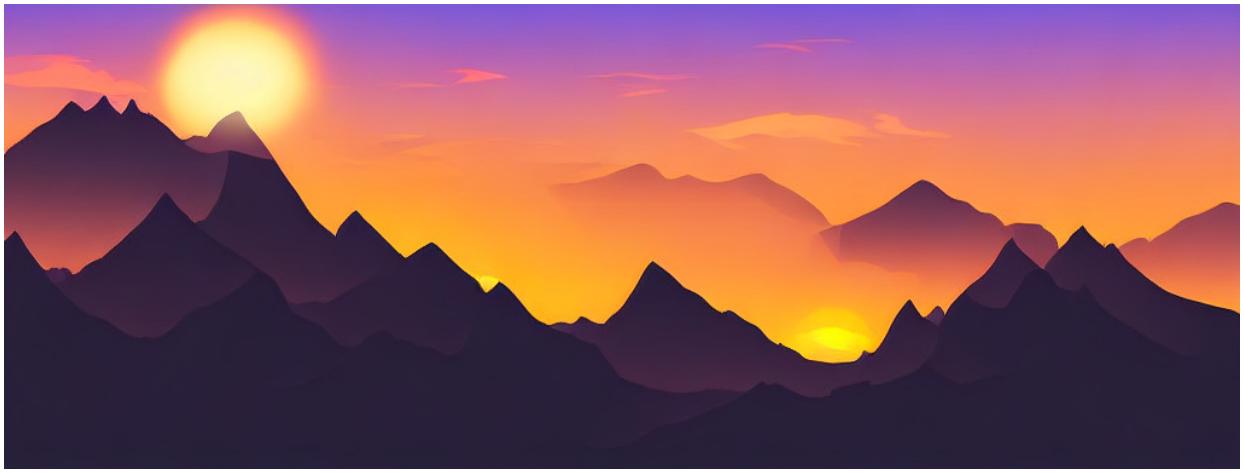

图13. 结合无分类器扩散引导与第4.3.2节的卷积采样策略，我们拥有14.5亿参数的文生图模型可用于渲染超出训练时原生256^2分辨率的图像。

A. Changelog

Here we list changes between this version (https://arxiv.org/abs/2112.10752v2) of the paper and the previous version, *i.e.* https://arxiv.org/abs/2112.10752v1.

- We updated the results on text-to-image synthesis in Sec. 4.3 which were obtained by training a new, larger model (1.45B parameters). This also includes a new comparison to very recent competing methods on this task that were published on arXiv at the same time as ([59, 109]) or after ([26]) the publication of our work.
- We updated results on class-conditional synthesis on ImageNet in Sec. 4.1, Tab. 3 (see also Sec. D.4) obtained by retraining the model with a larger batch size. The corresponding qualitative results in Fig. 26 and Fig. 27 were also updated. Both the updated text-to-image and the class-conditional model now use classifier-free guidance [32] as a measure to increase visual fidelity.
- We conducted a user study (following the scheme suggested by Saharia et al [72]) which provides additional evaluation for our inpainting (Sec. 4.5) and superresolution models (Sec. 4.4).
- Added Fig. 5 to the main paper, moved Fig. 18 to the appendix, added Fig. 13 to the appendix.

B. Detailed Information on Denoising Diffusion Models

Diffusion models can be specified in terms of a signal-to-noise ratio $\text{SNR}(t) = \frac{\alpha_t^2}{\sigma_t^2}$ consisting of sequences $(\alpha_t)_{t=1}^T$ and $(\sigma_t)_{t=1}^T$ which, starting from a data sample x_0, define a forward diffusion process q as

$$q(x_t|x_0) = \mathcal{N}(x_t|\alpha_t x_0, \sigma_t^2 \mathbb{I}) \tag{4}$$

with the Markov structure for $s < t$:

$$q(x_t|x_s) = \mathcal{N}(x_t|\alpha_{t|s} x_s, \sigma_{t|s}^2 \mathbb{I}) \tag{5}$$

$$\alpha_{t|s} = \frac{\alpha_t}{\alpha_s} \tag{6}$$

$$\sigma_{t|s}^2 = \sigma_t^2 - \alpha_{t|s}^2 \sigma_s^2 \tag{7}$$

Denoising diffusion models are generative models $p(x_0)$ which revert this process with a similar Markov structure running backward in time, *i.e.* they are specified as

$$p(x_0) = \int_z p(x_T) \prod_{t=1}^{T} p(x_{t-1}|x_t) \tag{8}$$

The evidence lower bound (ELBO) associated with this model then decomposes over the discrete time steps as

$$-\log p(x_0) \leq \mathbb{KL}(q(x_T|x_0)|p(x_T)) + \sum_{t=1}^{T} \mathbb{E}_{q(x_t|x_0)} \mathbb{KL}(q(x_{t-1}|x_t, x_0)|p(x_{t-1}|x_t)) \tag{9}$$

The prior $p(x_T)$ is typically choosen as a standard normal distribution and the first term of the ELBO then depends only on the final signal-to-noise ratio $\text{SNR}(T)$. To minimize the remaining terms, a common choice to parameterize $p(x_{t-1}|x_t)$ is to specify it in terms of the true posterior $q(x_{t-1}|x_t, x_0)$ but with the unknown x_0 replaced by an estimate $x_\theta(x_t, t)$ based on the current step x_t. This gives [45]

$$p(x_{t-1}|x_t) \coloneqq q(x_{t-1}|x_t, x_\theta(x_t, t)) \tag{10}$$

$$= \mathcal{N}(x_{t-1}|\mu_\theta(x_t, t), \sigma_{t|t-1}^2 \frac{\sigma_{t-1}^2}{\sigma_t^2} \mathbb{I}), \tag{11}$$

where the mean can be expressed as

$$\mu_\theta(x_t, t) = \frac{\alpha_{t|t-1}\sigma_{t-1}^2}{\sigma_t^2} x_t + \frac{\alpha_{t-1}\sigma_{t|t-1}^2}{\sigma_t^2} x_\theta(x_t, t). \tag{12}$$

A. 更新日志

此处列出本文当前版本（https://arxiv.org/abs/2112.10752v2）与先前版本*i.e*（https://arxiv.org/abs/2112.10752v1）之间的变更内容。

- 我们在第4.3节更新了文本到图像合成的结果，这些结果是通过训练一个更大规模的新模型（14.5亿参数）获得的。这还包括与近期同类方法的新比较，这些方法在我们工作发表的同时（[59, 109]）或之后（[26]）发布于arXiv。
- 我们在第4.1节和表3（另见第D.4节）中更新了ImageNet上类别条件合成的结果，这些结果是通过使用更大批量重新训练模型获得的。图26和图27中相应的定性结果也已更新。更新后的文本到图像模型和类别条件模型现在都采用无分类器引导{v*}作为提高视觉保真度的措施。
- 我们进行了一项用户研究（遵循Saharia等人[72]提出的方案），为我们的修复模型（第4.5节）和超分辨率模型（第4.4节）提供了额外评估。
- 在主论文中增加了图5，将图18移至附录，并在附录中增加了图13。

B. 去噪扩散模型详细信息

扩散模型可以通过信噪比SNR(t) $= \frac{\alpha_t^2}{\sigma_t^2}$来定义，该信噪比由序列$(\alpha_t)_{t=1}^T$和$(\sigma_t)_{t=1}^T$组成。从数据样本$x_0$出发，这些序列定义了一个前向扩散过程$q$，其形式为：

$$q(x_t|x_0) = \mathcal{N}(x_t|\alpha_t x_0, \sigma_t^2 \mathbb{I}) \tag{4}$$

在马尔可夫结构下，对于 $s < t$：

$$q(x_t|x_s) = \mathcal{N}(x_t|\alpha_{t|s} x_s, \sigma_{t|s}^2 \mathbb{I}) \tag{5}$$

$$\alpha_{t|s} = \frac{\alpha_t}{\alpha_s} \tag{6}$$

$$\sigma_{t|s}^2 = \sigma_t^2 - \alpha_{t|s}^2 \sigma_s^2 \tag{7}$$

去噪扩散模型是一种生成模型 $p(x_0)$，它通过类似马尔可夫结构在时间上反向运行来逆转这一过程 *i.e*，其具体定义为

$$p(x_0) = \int_z p(x_T) \prod_{t=1}^{T} p(x_{t-1}|x_t) \tag{8}$$

该模型对应的证据下界（ELBO）随后可分解为离散时间步上的求和形式：

$$-\log p(x_0) \leq \mathbb{KL}(q(x_T|x_0)|p(x_T)) + \sum_{t=1}^{T} \mathbb{E}_{q(x_t|x_0)} \mathbb{KL}(q(x_{t-1}|x_t, x_0)|p(x_{t-1}|x_t)) \tag{9}$$

先验 $p(x_T)$ 通常选择为标准正态分布，此时证据下界（ELBO）的第一项仅取决于最终的信噪比 SNR(*T*)。为最小化其余项，参数化 $p(x_{t-1}|x_t)$ 的常见方法是依据真实后验 $q(x_{t-1}|x_t, x_0)$ 进行设定，但将其中的未知量 x_0 替换为基于当前步骤 x_t 的估计值 $x_\theta(x_t, t)$。由此可得 [45]

$$p(x_{t-1}|x_t) \coloneqq q(x_{t-1}|x_t, x_\theta(x_t, t)) \tag{10}$$

$$= \mathcal{N}(x_{t-1}|\mu_\theta(x_t, t), \sigma_{t|t-1}^2 \frac{\sigma_{t-1}^2}{\sigma_t^2} \mathbb{I}), \tag{11}$$

其中均值可以表示为

$$\mu_\theta(x_t, t) = \frac{\alpha_{t|t-1}\sigma_{t-1}^2}{\sigma_t^2} x_t + \frac{\alpha_{t-1}\sigma_{t|t-1}^2}{\sigma_t^2} x_\theta(x_t, t). \tag{12}$$

In this case, the sum of the ELBO simplify to

$$\sum_{t=1}^{T} \mathbb{E}_{q(x_t|x_0)} \mathbb{KL}(q(x_{t-1}|x_t, x_0)|p(x_{t-1}) = \sum_{t=1}^{T} \mathbb{E}_{\mathcal{N}(\epsilon|0,\mathbb{I})} \frac{1}{2} (\text{SNR}(t-1) - \text{SNR}(t)) \|x_0 - x_\theta(\alpha_t x_0 + \sigma_t \epsilon, t)\|^2 \tag{13}$$

Following [30], we use the reparameterization

$$\epsilon_\theta(x_t, t) = (x_t - \alpha_t x_\theta(x_t, t))/\sigma_t \tag{14}$$

to express the reconstruction term as a denoising objective,

$$\|x_0 - x_\theta(\alpha_t x_0 + \sigma_t \epsilon, t)\|^2 = \frac{\sigma_t^2}{\alpha_t^2} \|\epsilon - \epsilon_\theta(\alpha_t x_0 + \sigma_t \epsilon, t)\|^2 \tag{15}$$

and the reweighting, which assigns each of the terms the same weight and results in Eq. (1).

在这种情况下，证据下界（ELBO）的总和简化为

$$\sum_{t=1}^{T}\mathbb{E}_{q(x_t|x_0)}\mathbb{KL}(q(x_{t-1}|x_t,x_0)|p(x_{t-1})=\sum_{t=1}^{T}\mathbb{E}_{\mathcal{N}(\epsilon|0,\mathbb{I})}\frac{1}{2}(\text{SNR}(t-1)-\text{SNR}(t))\|x_0-x_\theta(\alpha_t x_0+\sigma_t\epsilon,t)\|2 \tag{13}$$

根据[30]，我们使用重参数化

$$\epsilon_\theta(x_t,t)=(x_t-\alpha_t x_\theta(x_t,t))/\sigma_t \tag{14}$$

将重建项表达为去噪目标，

$$\|x_0-x_\theta(\alpha_t x_0+\sigma_t\epsilon,t)\|^2=\frac{\sigma_t^2}{\alpha_t^2}\|\epsilon-\epsilon_\theta(\alpha_t x_0+\sigma_t\epsilon,t)\|^2 \tag{15}$$

以及重新加权，它为每个项分配相同的权重，并得到公式(1)中的结果。

C. Image Guiding Mechanisms

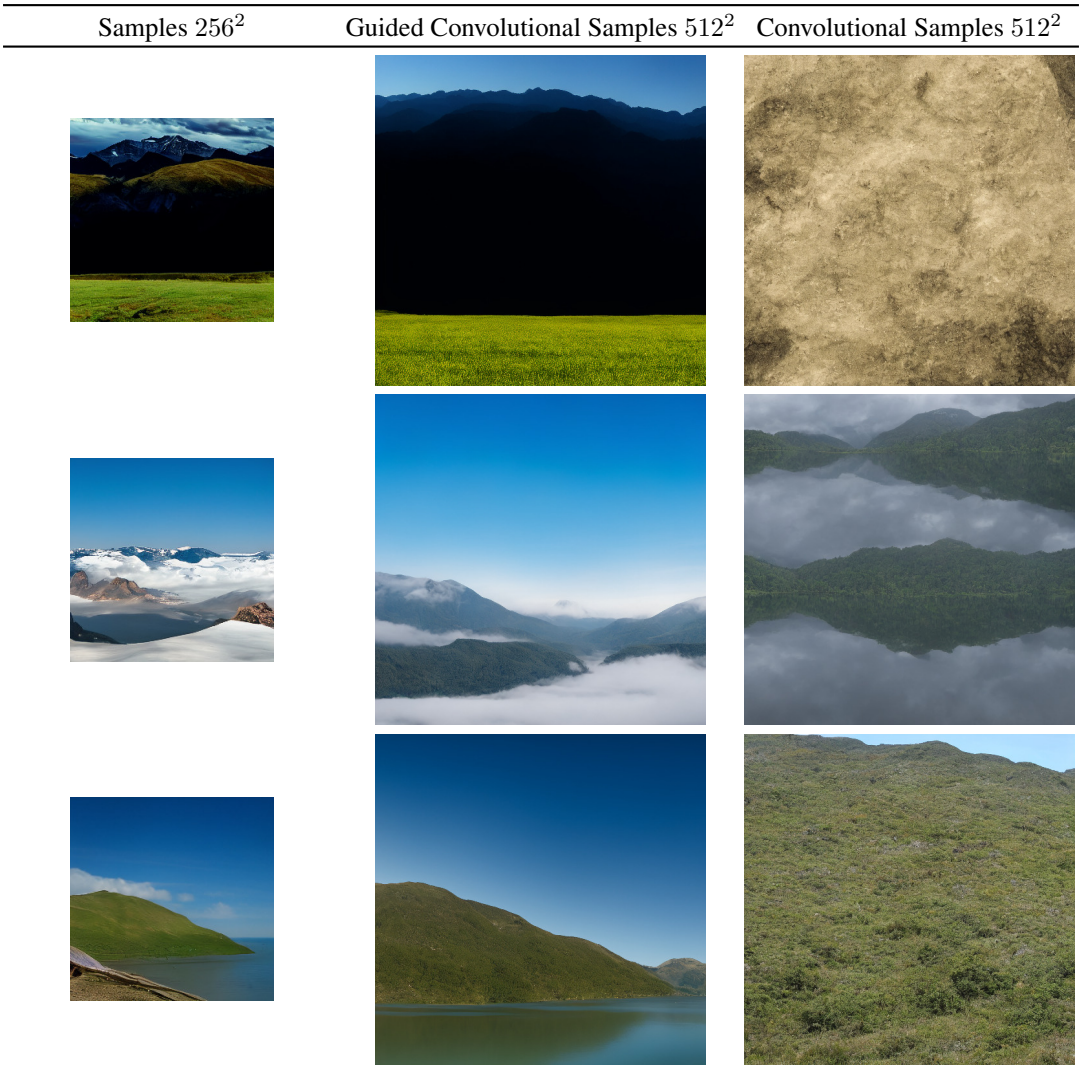

Figure 14. On landscapes, convolutional sampling with unconditional models can lead to homogeneous and incoherent global structures (see column 2). L_2-guiding with a low resolution image can help to reestablish coherent global structures.

An intriguing feature of diffusion models is that unconditional models can be conditioned at test-time [15, 82, 85]. In particular, [15] presented an algorithm to guide both unconditional and conditional models trained on the ImageNet dataset with a classifier $\log p_\Phi(y|x_t)$, trained on each x_t of the diffusion process. We directly build on this formulation and introduce post-hoc *image-guiding*:

For an epsilon-parameterized model with fixed variance, the guiding algorithm as introduced in [15] reads:

$$\hat{\epsilon} \leftarrow \epsilon_\theta(z_t, t) + \sqrt{1-\alpha_t^2}\, \nabla_{z_t} \log p_\Phi(y|z_t)\ . \tag{16}$$

This can be interpreted as an update correcting the "score" ϵ_θ with a conditional distribution $\log p_\Phi(y|z_t)$.

So far, this scenario has only been applied to single-class classification models. We re-interpret the guiding distribution $p_\Phi(y|T(\mathcal{D}(z_0(z_t))))$ as a general purpose image-to-image translation task given a target image y, where T can be any differentiable transformation adopted to the image-to-image translation task at hand, such as the identity, a downsampling operation or similar.

C. Image Guiding Mechanisms

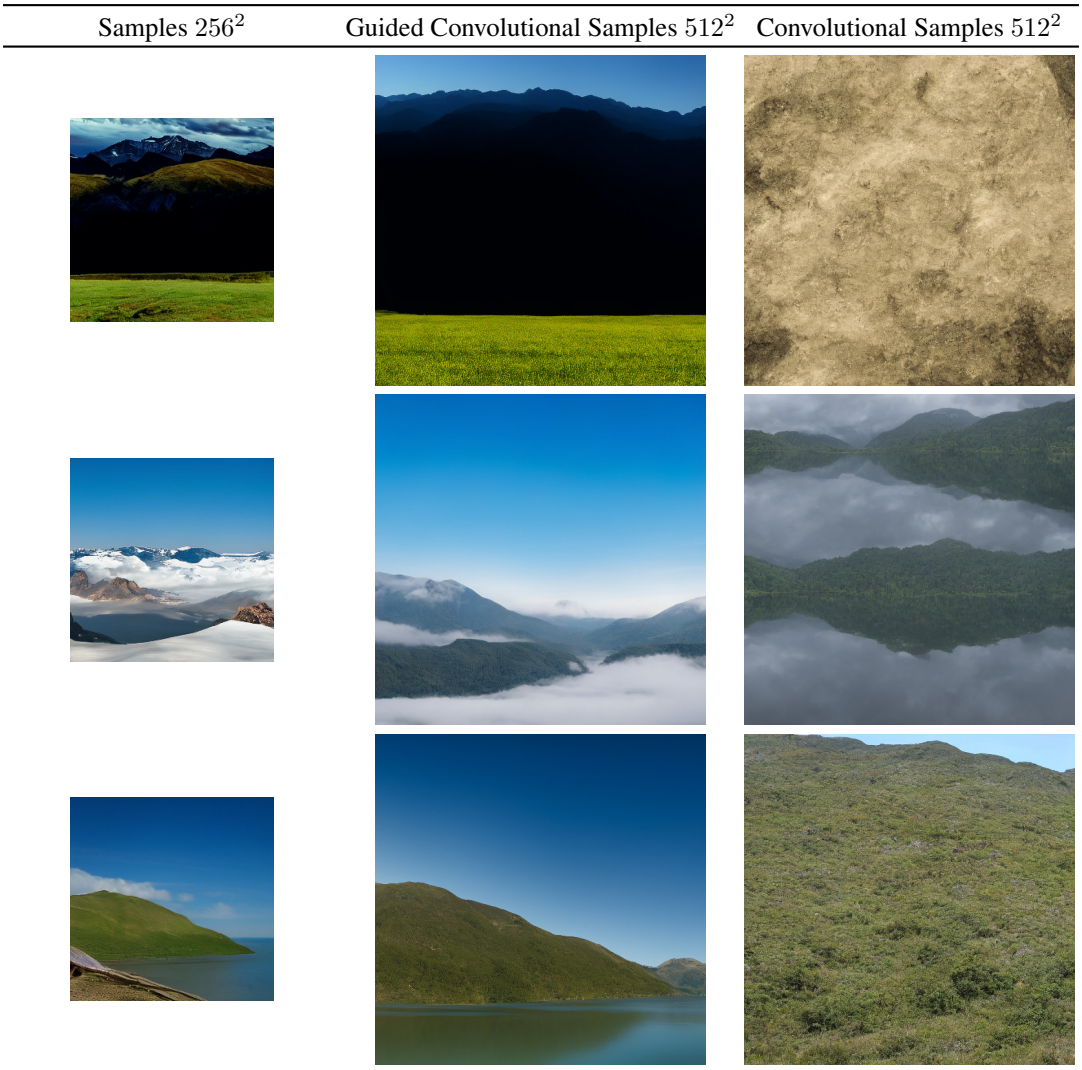

图14. 在风景场景中，无条件模型的卷积采样可能导致同质化且不连贯的全局结构（见第二列）。借助低分辨率图像进行L_2引导有助于重建连贯的全局结构。

扩散模型的一个有趣特性是，无条件的模型可以在测试时被条件化[15, 82, 85]。特别是，[15]提出了一种算法，用分类器对数$p_\Phi(y|x_t)$来引导在ImageNet数据集上训练的无条件和条件模型，该分类器在扩散过程的每个x_t上进行训练。我们直接基于这一表述，并引入事后*image-guiding*：

对于一个具有固定方差的epsilon参数化模型，[15]中介绍的引导算法表述为：

$$\hat{\epsilon} \leftarrow \epsilon_\theta(z_t, t) + \sqrt{1-\alpha_t^2}\, \nabla_{z_t} \log p_\Phi(y|z_t)\,. \tag{16}$$

这可以解释为 n 向丗期通过条件分布修正“分数” ϵ_θ g $p_\Phi(y|z_t)$。

迄今为止，这一方案仅应用于单类别分类模型。我们将引导分布$p_\Phi(y|T(\mathcal{D}(z_0(z_t))))$重新解释为给定目标图像$y$的通用图像到图像转换任务，其中$T$可以是适用于当前图像到图像转换任务的任何可微分变换，例如恒等变换、下采样操作或类似变换。

As an example, we can assume a Gaussian guider with fixed variance $\sigma^2 = 1$, such that

$$\log p_\Phi(y|z_t) = -\frac{1}{2}\|y - T(\mathcal{D}(z_0(z_t)))\|_2^2 \tag{17}$$

becomes a L_2 regression objective.

Fig. 14 demonstrates how this formulation can serve as an upsampling mechanism of an unconditional model trained on 256^2 images, where unconditional samples of size 256^2 guide the convolutional synthesis of 512^2 images and T is a $2\times$ bicubic downsampling. Following this motivation, we also experiment with a perceptual similarity guiding and replace the L_2 objective with the LPIPS [106] metric, see Sec. 4.4.

例如，我们可以假设一个具有固定方差 $\sigma^2 =$ 的高斯引导器，使得

$$\log p_\Phi(y|z_t) = -\frac{1}{2}\|y - T(\mathcal{D}(z_0(z_t)))\|_2^2 \tag{17}$$

成为一个 L_2 回归目标。

图14展示了该公式如何作为在256^2图像上训练的无条件模型的上采样机制，其中尺寸为256^2的无条件样本引导512^2图像的卷积合成，而T是2×双三次下采样。基于此动机，我们还尝试了感知相似性引导，并将L_2目标替换为LPIPS[106]度量，详见第4.4节。

D. Additional Results

D.1. Choosing the Signal-to-Noise Ratio for High-Resolution Synthesis

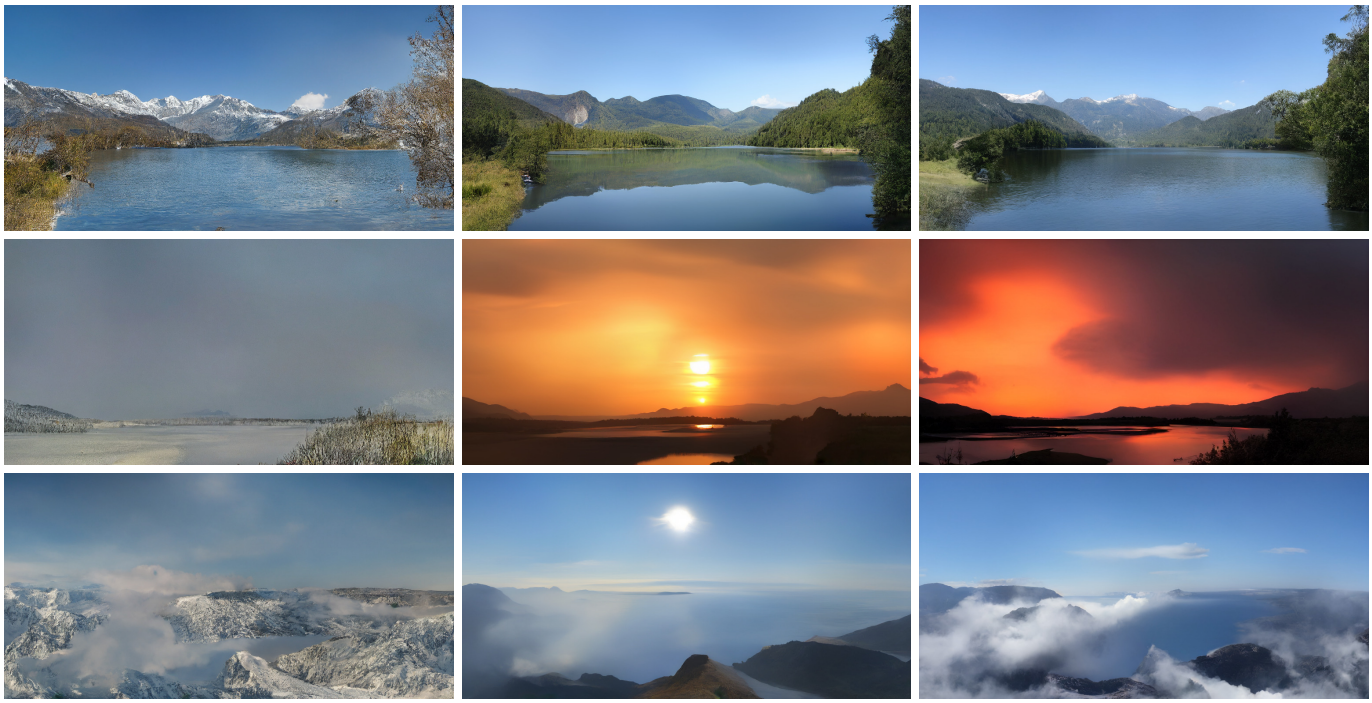

Figure 15. Illustrating the effect of latent space rescaling on convolutional sampling, here for semantic image synthesis on landscapes. See Sec. 4.3.2 and Sec. D.1.

As discussed in Sec. 4.3.2, the signal-to-noise ratio induced by the variance of the latent space (*i.e.* $\mathrm{Var}(z)/\sigma_t^2$) significantly affects the results for convolutional sampling. For example, when training a LDM directly in the latent space of a KL-regularized model (see Tab. 8), this ratio is very high, such that the model allocates a lot of semantic detail early on in the reverse denoising process. In contrast, when rescaling the latent space by the component-wise standard deviation of the latents as described in Sec. G, the SNR is descreased. We illustrate the effect on convolutional sampling for semantic image synthesis in Fig. 15. Note that the VQ-regularized space has a variance close to 1, such that it does not have to be rescaled.

D.2. Full List of all First Stage Models

We provide a complete list of various autoenconding models trained on the OpenImages dataset in Tab. 8.

D.3. Layout-to-Image Synthesis

Here we provide the quantitative evaluation and additional samples for our layout-to-image models from Sec. 4.3.1. We train a model on the COCO [4] and one on the OpenImages [49] dataset, which we subsequently additionally finetune on COCO. Tab 9 shows the result. Our COCO model reaches the performance of recent state-of-the art models in layout-to-image synthesis, when following their training and evaluation protocol [89]. When finetuning from the OpenImages model, we surpass these works. Our OpenImages model surpasses the results of Jahn et al [37] by a margin of nearly 11 in terms of FID. In Fig. 16 we show additional samples of the model finetuned on COCO.

D.4. Class-Conditional Image Synthesis on ImageNet

Tab. 10 contains the results for our class-conditional LDM measured in FID and Inception score (IS). LDM-8 requires significantly fewer parameters and compute requirements (see Tab. 18) to achieve very competitive performance. Similar to previous work, we can further boost the performance by training a classifier on each noise scale and guiding with it,

D. 补充结果

D.1. 高分辨率合成中信号噪声比的选择

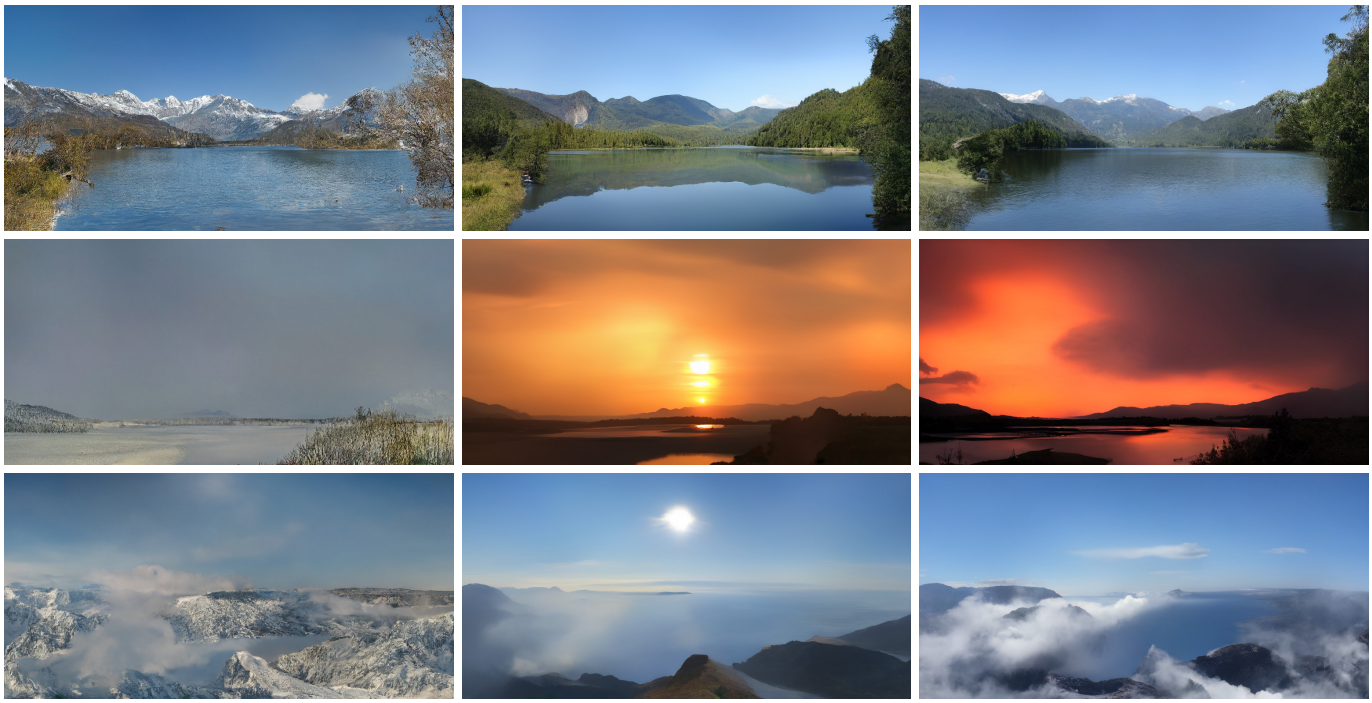

图15. 展示了潜在空间重缩放对卷积采样的影响，此处以景观语义图像合成为例。详见第4.3.2节与第D.1节。

如第4.3.2节所述，潜在空间方差（*i.e.* $\mathrm{Var}(z)/\sigma_t^2$）引起的信噪比显著影响卷积采样的结果。例如，在KL正则化模型的潜在空间中直接训练LDM时（见表8），该比率非常高，导致模型在反向去噪过程的早期就分配了大量语义细节。相反，如第G节所述，通过按潜在分量的标准差重新缩放潜在空间时，信噪比会降低。我们在图15中展示了卷积采样对语义图像合成的影响。需要注意的是，VQ正则化空间的方差接近1，因此无需重新缩放。

D.2. 第一阶段模型完整列表

我们在表8中提供了在OpenImages数据集上训练的各种自动编码模型的完整列表。

D.3. 布局到图像合成

在此，我们提供了第4.3.1节中布局到图像模型的定量评估及额外样本。我们在COCO数据集[4]上训练了一个模型，并在OpenImages数据集[49]上训练了另一个模型，随后又在COCO上对后者进行了微调。表9展示了相关结果。在遵循现有研究[89]的训练与评估协议时，我们的COCO模型达到了近期布局到图像合成领域先进模型的性能水平。而通过对OpenImages模型进行微调，我们进一步超越了这些研究成果。在FID指标上，我们的OpenImages模型以近11分的优势超越了Jahn等人[37]的结果。图16展示了在COCO上微调后模型的更多生成样本。

D.4. ImageNet上的类别条件图像合成

表10展示了我们按类别条件化的LDM在FID和初始分数（IS）上的评估结果。LDM-8仅需显著更少的参数量与计算需求（见表18）即可实现极具竞争力的性能。与先前工作类似，我们还能通过在每个噪声尺度上训练分类器并进行引导来进一步提升性能，

f	$\lvert\mathcal{Z}\rvert$	c	**R-FID** ↓	**R-IS** ↑	**PSNR** ↑	**PSIM** ↓	**SSIM** ↑
16 *VQGAN* [23]	16384	256	4.98	–	19.9 ±3.4	1.83 ±0.42	0.51 ±0.18
16 *VQGAN* [23]	1024	256	7.94	–	19.4 ±3.3	1.98 ±0.43	0.50 ±0.18
8 *DALL-E* [66]	8192	-	32.01	–	22.8 ±2.1	1.95 ±0.51	0.73 ±0.13
32	16384	16	31.83	40.40 ±1.07	17.45 ±2.90	2.58 ±0.48	0.41 ±0.18
16	16384	8	5.15	144.55 ±3.74	20.83 ±3.61	1.73 ±0.43	0.54 ±0.18
8	16384	4	1.14	201.92 ±3.97	23.07 ±3.99	1.17 ±0.36	0.65 ±0.16
8	256	4	1.49	194.20 ±3.87	22.35 ±3.81	1.26 ±0.37	0.62 ±0.16
4	8192	3	0.58	224.78 ±5.35	27.43 ±4.26	0.53 ±0.21	0.82 ±0.10
4†	8192	3	1.06	221.94 ±4.58	25.21 ±4.17	0.72 ±0.26	0.76 ±0.12
4	256	3	0.47	223.81 ±4.58	26.43 ±4.22	0.62 ±0.24	0.80 ±0.11
2	2048	2	0.16	232.75 ±5.09	30.85 ±4.12	0.27 ±0.12	0.91 ±0.05
2	64	2	0.40	226.62 ±4.83	29.13 ±3.46	0.38 ±0.13	0.90 ±0.05
32	KL	64	2.04	189.53 ±3.68	22.27 ±3.93	1.41 ±0.40	0.61 ±0.17
32	KL	16	7.3	132.75 ±2.71	20.38 ±3.56	1.88 ±0.45	0.53 ±0.18
16	KL	16	0.87	210.31 ±3.97	24.08 ±4.22	1.07 ±0.36	0.68 ±0.15
16	KL	8	2.63	178.68 ±4.08	21.94 ±3.92	1.49 ±0.42	0.59 ±0.17
8	KL	4	0.90	209.90 ±4.92	24.19 ±4.19	1.02 ±0.35	0.69 ±0.15
4	KL	3	0.27	227.57 ±4.89	27.53 ±4.54	0.55 ±0.24	0.82 ±0.11
2	KL	2	0.086	232.66 ±5.16	32.47 ±4.19	0.20 ±0.09	0.93 ±0.04

Table 8. Complete autoencoder zoo trained on OpenImages, evaluated on ImageNet-Val. † denotes an attention-free autoencoder.

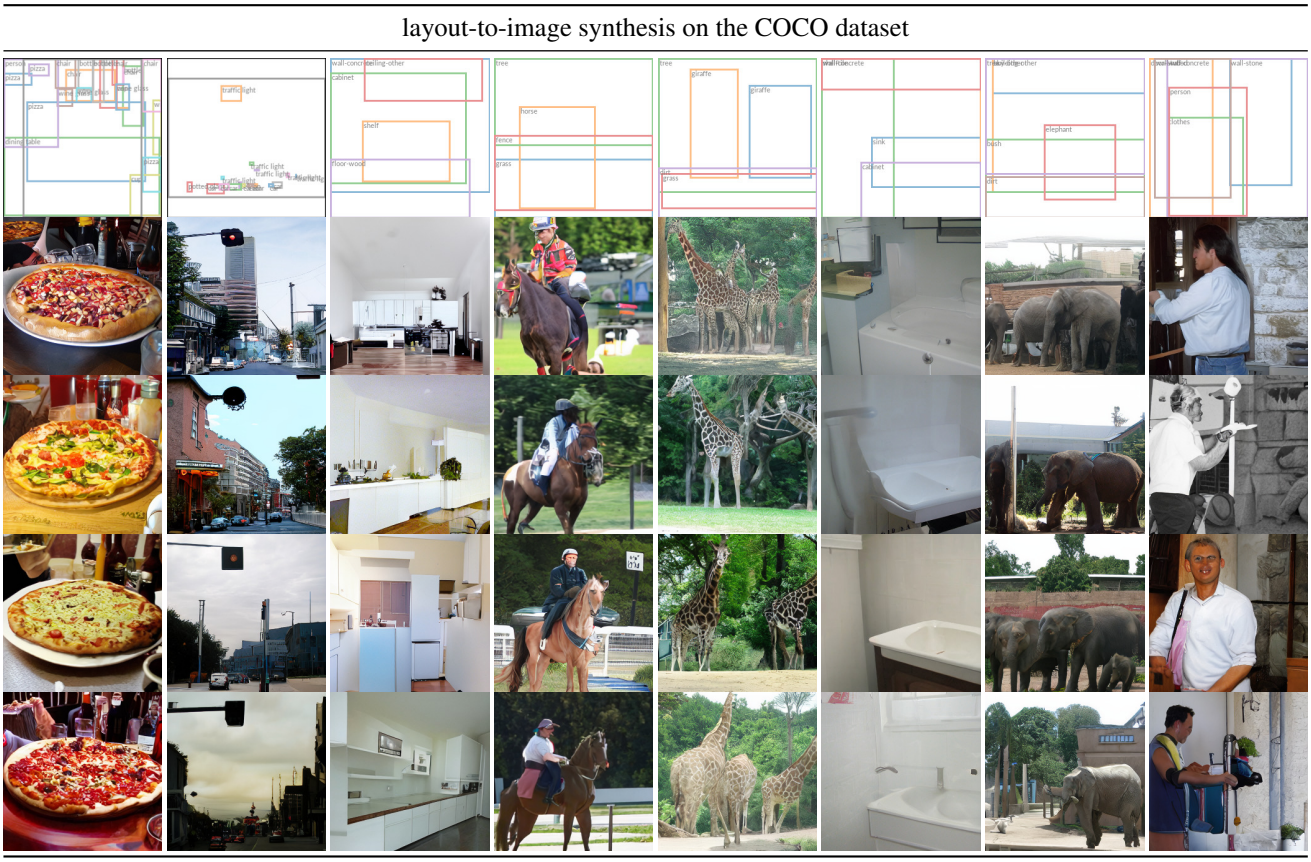

Figure 16. More samples from our best model for layout-to-image synthesis, *LDM-4*, which was trained on the OpenImages dataset and finetuned on the COCO dataset. Samples generated with 100 DDIM steps and $\eta = 0$. Layouts are from the COCO validation set.

see Sec. C. Unlike the pixel-based methods, this classifier is trained very cheaply in latent space. For additional qualitative results, see Fig. 26 and Fig. 27.

f	$\lvert\mathcal{Z}\rvert$	c	R-FID ↓	R-IS ↑	PSNR ↑	PSIM ↓	SSIM ↑
16 *VQGAN* [23]	16384	256	4.98	–	19.9 ±3.4	1.83 ±0.42	0.51 ±0.18
16 *VQGAN* [23]	1024	256	7.94	–	19.4 ±3.3	1.98 ±0.43	0.50 ±0.18
8 *DALL-E* [66]	8192	-	32.01	–	22.8 ±2.1	1.95 ±0.51	0.73 ±0.13
32	16384	16	31.83	40.40 ±1.07	17.45 ±2.90	2.58 ±0.48	0.41 ±0.18
16	16384	8	5.15	144.55 ±3.74	20.83 ±3.61	1.73 ±0.43	0.54 ±0.18
8	16384	4	1.14	201.92 ±3.97	23.07 ±3.99	1.17 ±0.36	0.65 ±0.16
8	256	4	1.49	194.20 ±3.87	22.35 ±3.81	1.26 ±0.37	0.62 ±0.16
4	8192	3	0.58	224.78 ±5.35	27.43 ±4.26	0.53 ±0.21	0.82 ±0.10
4†	8192	3	1.06	221.94 ±4.58	25.21 ±4.17	0.72 ±0.26	0.76 ±0.12
4	256	3	0.47	223.81 ±4.58	26.43 ±4.22	0.62 ±0.24	0.80 ±0.11
2	2048	2	0.16	232.75 ±5.09	30.85 ±4.12	0.27 ±0.12	0.91 ±0.05
2	64	2	0.40	226.62 ±4.83	29.13 ±3.46	0.38 ±0.13	0.90 ±0.05
32	KL	64	2.04	189.53 ±3.68	22.27 ±3.93	1.41 ±0.40	0.61 ±0.17
32	KL	16	7.3	132.75 ±2.71	20.38 ±3.56	1.88 ±0.45	0.53 ±0.18
16	KL	16	0.87	210.31 ±3.97	24.08 ±4.22	1.07 ±0.36	0.68 ±0.15
16	KL	8	2.63	178.68 ±4.08	21.94 ±3.92	1.49 ±0.42	0.59 ±0.17
8	KL	4	0.90	209.90 ±4.92	24.19 ±4.19	1.02 ±0.35	0.69 ±0.15
4	KL	3	0.27	227.57 ±4.89	27.53 ±4.54	0.55 ±0.24	0.82 ±0.11
2	KL	2	0.086	232.66 ±5.16	32.47 ±4.19	0.20 ±0.09	0.93 ±0.04

表8. 完整自动编码 在OpenImages上训练，在ImageNet-Val上评估。†表示 一种无注意力的自编码器。

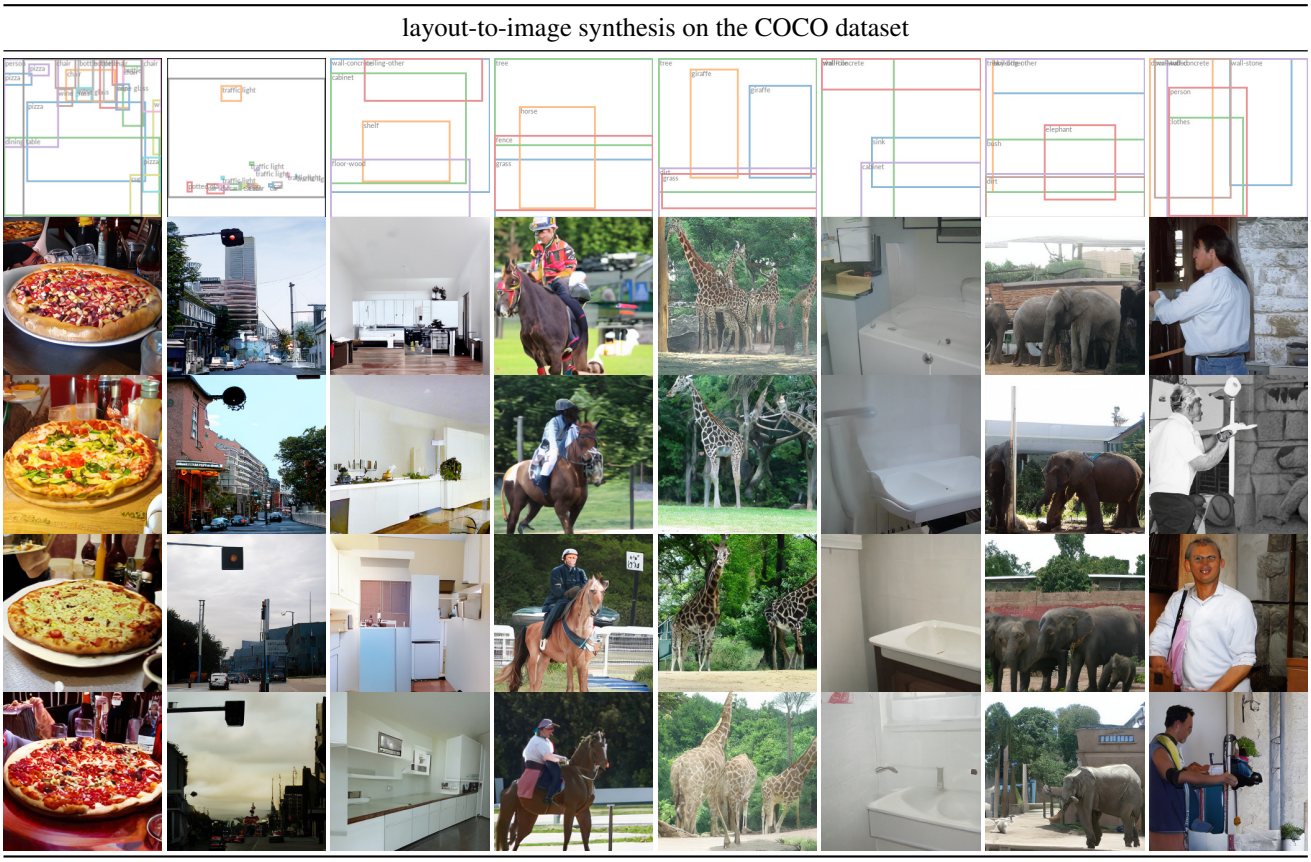

图16. 来自我们最佳布局到图像合成模型 {v*} 的更多样本，该模型在OpenImages数据集上训练并在COCO数据集上微调。样本使用100步DDIM和 {v*} 0生成。布局来自COCO验证集。

参见附录C。与基于像素的方法不同，该分类器在潜在空间中的训练成本极低。更多定性结果请参见图26和图27。

	COCO256 × 256	OpenImages 256 × 256	OpenImages 512 × 512
Method	FID↓	FID↓	FID↓
LostGAN-V2 [87]	42.55	-	-
OC-GAN [89]	41.65	-	-
SPADE [62]	41.11	-	-
VQGAN+T [37]	56.58	45.33	48.11
LDM-8 (100 steps, ours)	42.06†	-	-
LDM-4 (200 steps, ours)	**40.91***	**32.02**	**35.80**

Table 9. Quantitative comparison of our layout-to-image models on the COCO [4] and OpenImages [49] datasets. †: Training from scratch on COCO; *: Finetuning from OpenImages.

Method	FID↓	IS↑	Precision↑	Recall↑	N_{params}	
SR3 [72]	11.30	-	-	-	625M	-
ImageBART [21]	21.19	-	-	-	3.5B	-
ImageBART [21]	7.44	-	-	-	3.5B	0.05 acc. rate*
VQGAN+T [23]	17.04	70.6±1.8	-	-	1.3B	-
VQGAN+T [23]	5.88	**304.8**±3.6	-	-	1.3B	0.05 acc. rate*
BigGan-deep [3]	6.95	203.6±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
ADM-G,ADM-U [15]	3.85	221.72	0.84	0.53	n/a	2 × 250 DDIM steps
CDM [31]	4.88	158.71±2.26	-	-	n/a	2 × 100 DDIM steps
LDM-8 (ours)	17.41	72.92±2.6	0.65	0.62	395M	200 DDIM steps, 2.9M train steps, batch size 64
LDM-8-G (ours)	8.11	190.43±2.60	0.83	0.36	506M	200 DDIM steps, classifier scale 10, 2.9M train steps, batch size 64
LDM-8 (ours)	15.51	79.03±1.03	0.65	**0.63**	395M	200 DDIM steps, 4.8M train steps, batch size 64
LDM-8-G (ours)	7.76	209.52±4.24	0.84	0.35	506M	200 DDIM steps, classifier scale 10, 4.8M train steps, batch size 64
LDM-4 (ours)	10.56	103.49±1.24	0.71	0.62	400M	250 DDIM steps, 178K train steps, batch size 1200
LDM-4-G (ours)	3.95	178.22±2.43	0.81	0.55	400M	250 DDIM steps, unconditional guidance [32] scale 1.25, 178K train steps, batch size 1200
LDM-4-G (ours)	**3.60**	247.67±5.59	**0.87**	0.48	400M	250 DDIM steps, unconditional guidance [32] scale 1.5, 178K train steps, batch size 1200

Table 10. Comparison of a class-conditional ImageNet *LDM* with recent state-of-the-art methods for class-conditional image generation on the ImageNet [12] dataset.*: Classifier rejection sampling with the given rejection rate as proposed in [67].

D.5. Sample Quality vs. V100 Days (Continued from Sec. 4.1)

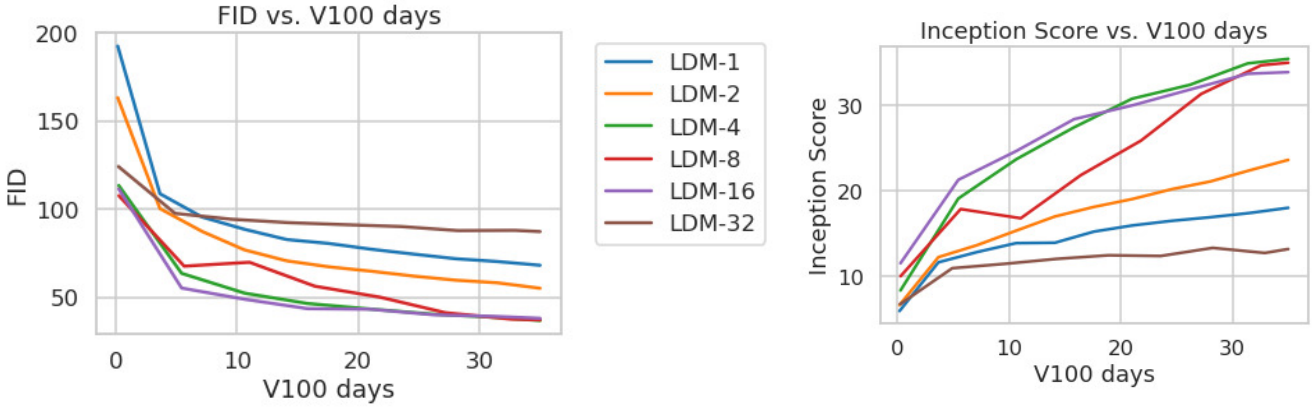

Figure 17. For completeness we also report the training progress of class-conditional *LDMs* on the ImageNet dataset for a fixed number of 35 V100 days. Results obtained with 100 DDIM steps [84] and $\kappa = 0$. FIDs computed on 5000 samples for efficiency reasons.

For the assessment of sample quality over the training progress in Sec. 4.1, we reported FID and IS scores as a function of train steps. Another possibility is to report these metrics over the used resources in V100 days. Such an analysis is additionally provided in Fig. 17, showing qualitatively similar results.

	COCO256 × 256	OpenImages 256 × 256	OpenImages 512 × 512
Method	FID↓	FID↓	FID↓
LostGAN-V2 [87]	42.55	-	-
OC-GAN [89]	41.65	-	-
SPADE [62]	41.11	-	-
VQGAN+T [37]	56.58	45.33	48.11
LDM-8 (100 steps, ours)	42.06†	-	-
LDM-4 (200 steps, ours)	**40.91***	**32.02**	**35.80**

表9. 我们的布局到图像模型在COCO [4]和OpenImages [49]数据集上的定量比较。†: 在COCO上从头开始训练；*: 从OpenImages进行微调。

Method	FID↓	IS↑	Precision↑	Recall↑	N_{params}	
SR3 [72]	11.30	-	-	-	625M	-
ImageBART [21]	21.19	-	-	-	3.5B	-
ImageBART [21]	7.44	-	-	-	3.5B	0.05 acc. rate*
VQGAN+T [23]	17.04	70.6±1.8	-	-	1.3B	-
VQGAN+T [23]	5.88	**304.8**±3.6	-	-	1.3B	0.05 acc. rate*
BigGan-deep [3]	6.95	203.6±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
ADM-G,ADM-U [15]	3.85	221.72	0.84	0.53	n/a	2 × 250 DDIM steps
CDM [31]	4.88	158.71±2.26	-	-	n/a	2 × 100 DDIM steps
LDM-8 (ours)	17.41	72.92±2.6	0.65	0.62	395M	200 DDIM steps, 2.9M train steps, batch size 64
LDM-8-G (ours)	8.11	190.43±2.60	0.83	0.36	506M	200 DDIM steps, classifier scale 10, 2.9M train steps, batch size 64
LDM-8 (ours)	15.51	79.03±1.03	0.65	**0.63**	395M	200 DDIM steps, 4.8M train steps, batch size 64
LDM-8-G (ours)	7.76	209.52±4.24	0.84	0.35	506M	200 DDIM steps, classifier scale 10, 4.8M train steps, batch size 64
LDM-4 (ours)	10.56	103.49±1.24	0.71	0.62	400M	250 DDIM steps, 178K train steps, batch size 1200
LDM-4-G (ours)	3.95	178.22±2.43	0.81	0.55	400M	250 DDIM steps, unconditional guidance [32] scale 1.25, 178K train steps, batch size 1200
LDM-4-G (ours)	**3.60**	247.67±5.59	**0.87**	0.48	400M	250 DDIM steps, unconditional guidance [32] scale 1.5, 178K train steps, batch size 1200

表10. 在ImageNet [12]数据集上，类别条件ImageNet *LDM*与近期最先进的类别条件图像生成方法的比较。*：采用[67]中提出的给定拒绝率进行分类器拒绝采样。

D.5. 样本质量与V100天数对比（续4.1节）

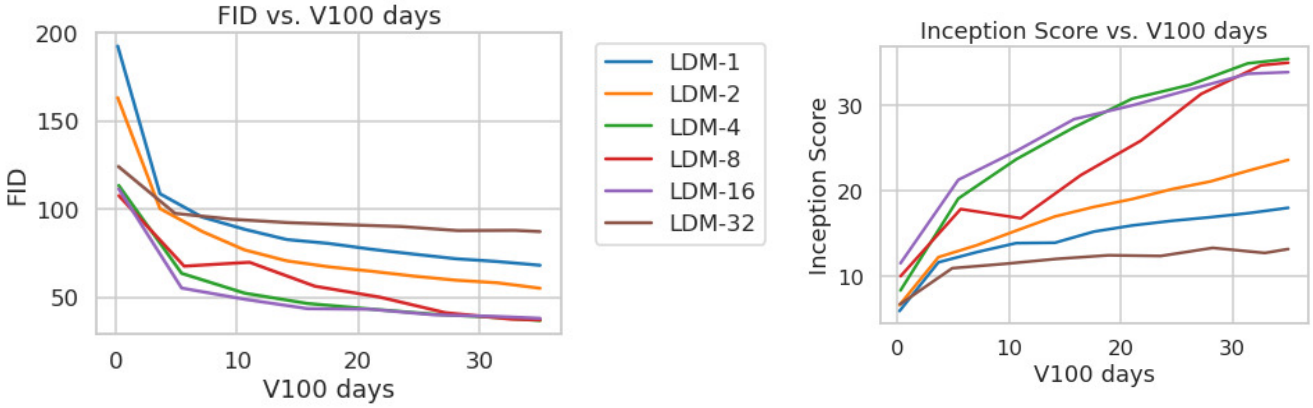

图17。为完整起见，我们还报告了类别条件*LDMs*在ImageNet数据集上以固定35个V100天计算的训练进度。结果采用100步DDIM采样[84]和$\kappa = 0$获得。为提升计算效率，FID指标基于5000个样本计算得出。

在4.1节中评估训练过程中的样本质量时，我们报告了FID和IS分数随训练步数的变化关系。另一种可能性是报告这些指标在V100天资源消耗下的表现。图17额外提供了此类分析，结果显示其定性结论基本一致。

Method	FID ↓	IS ↑	PSNR ↑	SSIM ↑
Image Regression [72]	15.2	121.1	**27.9**	**0.801**
SR3 [72]	5.2	**180.1**	26.4	0.762
LDM-4 (ours, 100 steps)	**2.8**†/**4.8**‡	166.3	24.4$\pm$3.8	0.69$\pm$0.14
LDM-4 (ours, 50 steps, guiding)	4.4†/6.4‡	153.7	25.8$\pm$3.7	0.74$\pm$0.12
LDM-4 (ours, 100 steps, guiding)	4.4†/6.4‡	154.1	25.7$\pm$3.7	0.73$\pm$0.12
LDM-4 (ours, 100 steps, +15 ep.)	**2.6**† / **4.6**‡	169.76$\pm$5.03	24.4$\pm$3.8	0.69$\pm$0.14
Pixel-DM (100 steps, +15 ep.)	5.1† / 7.1‡	163.06$\pm$4.67	24.1$\pm$3.3	0.59$\pm$0.12

Table 11. $\times 4$ upscaling results on ImageNet-Val. (256^2); †: FID features computed on validation split, ‡: FID features computed on train split. We also include a pixel-space baseline that receives the same amount of compute as *LDM-4*. The last two rows received 15 epochs of additional training compared to the former results.

D.6. Super-Resolution

For better comparability between LDMs and diffusion models in pixel space, we extend our analysis from Tab. 5 by comparing a diffusion model trained for the same number of steps and with a comparable number [1] of parameters to our LDM. The results of this comparison are shown in the last two rows of Tab. 11 and demonstrate that LDM achieves better performance while allowing for significantly faster sampling. A qualitative comparison is given in Fig. 20 which shows random samples from both LDM and the diffusion model in pixel space.

D.6.1 LDM-BSR: General Purpose SR Model via Diverse Image Degradation

Figure 18. *LDM-BSR* generalizes to arbitrary inputs and can be used as a general-purpose upsampler, upscaling samples from a class-conditional *LDM* (image *cf*. Fig. 4) to 1024^2 resolution. In contrast, using a fixed degradation process (see Sec. 4.4) hinders generalization.

To evaluate generalization of our LDM-SR, we apply it both on synthetic LDM samples from a class-conditional ImageNet model (Sec. 4.1) and images crawled from the internet. Interestingly, we observe that LDM-SR, trained only with a bicubicly downsampled conditioning as in [72], does not generalize well to images which do not follow this pre-processing. Hence, to obtain a superresolution model for a wide range of real world images, which can contain complex superpositions of camera noise, compression artifacts, blurr and interpolations, we replace the bicubic downsampling operation in LDM-SR with the degration pipeline from [105]. The BSR-degradation process is a degradation pipline which applies JPEG compressions noise, camera sensor noise, different image interpolations for downsampling, Gaussian blur kernels and Gaussian noise in a random order to an image. We found that using the bsr-degredation process with the original parameters as in [105] leads to a very strong degradation process. Since a more moderate degradation process seemed apppropiate for our application, we adapted the parameters of the bsr-degradation (our adapted degradation process can be found in our code base at `https://github.com/CompVis/latent-diffusion`). Fig. 18 illustrates the effectiveness of this approach by directly comparing *LDM-SR* with *LDM-BSR*. The latter produces images much sharper than the models confined to a fixed pre-processing, making it suitable for real-world applications. Further results of *LDM-BSR* are shown on LSUN-cows in Fig. 19.

[1] It is not possible to exactly match both architectures since the diffusion model operates in the pixel space

Method	FID ↓	IS ↑	PSNR ↑	SSIM ↑
Image Regression [72]	15.2	121.1	**27.9**	**0.801**
SR3 [72]	5.2	**180.1**	26.4	0.762
LDM-4 (ours, 100 steps)	**2.8**†/**4.8**‡	166.3	24.4±3.8	0.69±0.14
LDM-4 (ours, 50 steps, guiding)	4.4†/6.4‡	153.7	25.8±3.7	0.74±0.12
LDM-4 (ours, 100 steps, guiding)	4.4†/6.4‡	154.1	25.7±3.7	0.73±0.12
LDM-4 (ours, 100 steps, +15 ep.)	**2.6**† / **4.6**‡	169.76±5.03	24.4±3.8	0.69±0.14
Pixel-DM (100 steps, +15 ep.)	5.1† / 7.1‡	163.06±4.67	24.1±3.3	0.59±0.12

表 11. ImageNet-Val 上的 ×4 上采样结果（256^2）；†：基于验证集计算的 FID 特征，‡：基于训练集计算的 FID 特征。我们还纳入了一个与 *LDM-4* 计算量相同的像素空间基线。最后两行结果较先前增加了 15 个训练周期。

D.6. 超分辨率

为了在像素空间中更好地比较LDM与扩散模型，我们将表5的分析扩展，比较了经过相同训练步数且参数量[1]相当的扩散模型与我们的LDM。该比较结果展示在表11的最后两行，表明LDM在实现更优性能的同时允许显著更快的采样速度。图20提供了定性对比，展示了像素空间中LDM与扩散模型的随机生成样本。

D.6.1 LDM-BSR：通过多样化图像退化实现的通用超分辨率模型

图18。*LDM-BSR*可泛化至任意输入，并能作为通用上采样器使用，将来自类别条件*LDM* (图像*cf*的样本从图4)的分辨率提升至1024^2。相比之下，使用固定的退化过程（见第4.4节）会阻碍泛化能力。

为了评估我们的LDM-SR的泛化能力，我们将其应用于来自类别条件ImageNet模型的合成LDM样本（第4.1节）以及从互联网爬取的图像。有趣的是，我们观察到，仅使用如[72]中所述的双三次下采样条件进行训练的LDM-SR，对于不遵循此预处理的图像泛化效果不佳。因此，为了获得适用于广泛现实世界图像的超分辨率模型——这些图像可能包含相机噪声、压缩伪影、模糊和插值的复杂叠加——我们将LDM-SR中的双三次下采样操作替换为[105]中的退化流程。BSR退化过程是一个退化流程，它按随机顺序对图像应用JPEG压缩噪声、相机传感器噪声、用于下采样的不同图像插值、高斯模糊核和高斯噪声。我们发现，使用[105]中原始参数的bsr退化过程会导致非常强烈的退化效果。由于我们的应用似乎更适合一个更温和的退化过程，我们调整了bsr退化的参数（我们调整后的退化过程可在我们的代码库https://github.com/CompVis/latent-diffusion中找到）。图18通过直接比较*LDM-SR*与*LDM-BSR*说明了这种方法的有效性。后者生成的图像比局限于固定预处理的模型锐利得多，使其适用于实际应用。*LDM-BSR*的更多结果展示在图19的LSUN-cows数据集上。

[1] It is not possible to exactly match both architectures since the diffusion model operates in the pixel space

E. Implementation Details and Hyperparameters

E.1. Hyperparameters

We provide an overview of the hyperparameters of all trained *LDM* models in Tab. 12, Tab. 13, Tab. 14 and Tab. 15.

	CelebA-HQ 256×256	FFHQ 256×256	LSUN-Churches 256×256	LSUN-Bedrooms 256×256
f	4	4	8	4
z-shape	$64 \times 64 \times 3$	$64 \times 64 \times 3$	-	$64 \times 64 \times 3$
$\lvert\mathcal{Z}\rvert$	8192	8192	-	8192
Diffusion steps	1000	1000	1000	1000
Noise Schedule	linear	linear	linear	linear
N_{params}	274M	274M	294M	274M
Channels	224	224	192	224
Depth	2	2	2	2
Channel Multiplier	1,2,3,4	1,2,3,4	1,2,2,4,4	1,2,3,4
Attention resolutions	32, 16, 8	32, 16, 8	32, 16, 8, 4	32, 16, 8
Head Channels	32	32	24	32
Batch Size	48	42	96	48
Iterations*	410k	635k	500k	1.9M
Learning Rate	9.6e-5	8.4e-5	5.e-5	9.6e-5

Table 12. Hyperparameters for the unconditional *LDMs* producing the numbers shown in Tab. 1. All models trained on a single NVIDIA A100.

	LDM-1	*LDM-2*	*LDM-4*	*LDM-8*	*LDM-16*	*LDM-32*
z-shape	$256 \times 256 \times 3$	$128 \times 128 \times 2$	$64 \times 64 \times 3$	$32 \times 32 \times 4$	$16 \times 16 \times 8$	$88 \times 8 \times 32$
$\lvert\mathcal{Z}\rvert$	-	2048	8192	16384	16384	16384
Diffusion steps	1000	1000	1000	1000	1000	1000
Noise Schedule	linear	linear	linear	linear	linear	linear
Model Size	396M	391M	391M	395M	395M	395M
Channels	192	192	192	256	256	256
Depth	2	2	2	2	2	2
Channel Multiplier	1,1,2,2,4,4	1,2,2,4,4	1,2,3,5	1,2,4	1,2,4	1,2,4
Number of Heads	1	1	1	1	1	1
Batch Size	7	9	40	64	112	112
Iterations	2M	2M	2M	2M	2M	2M
Learning Rate	4.9e-5	6.3e-5	8e-5	6.4e-5	4.5e-5	4.5e-5
Conditioning	CA	CA	CA	CA	CA	CA
CA-resolutions	32, 16, 8	32, 16, 8	32, 16, 8	32, 16, 8	16, 8, 4	8, 4, 2
Embedding Dimension	512	512	512	512	512	512
Transformers Depth	1	1	1	1	1	1

Table 13. Hyperparameters for the conditional *LDMs* trained on the ImageNet dataset for the analysis in Sec. 4.1. All models trained on a single NVIDIA A100.

E.2. Implementation Details

E.2.1 Implementations of τ_θ for conditional *LDMs*

For the experiments on text-to-image and layout-to-image (Sec. 4.3.1) synthesis, we implement the conditioner τ_θ as an unmasked transformer which processes a tokenized version of the input y and produces an output $\zeta := \tau_\theta(y)$, where $\zeta \in \mathbb{R}^{M \times d_\tau}$. More specifically, the transformer is implemented from N transformer blocks consisting of global self-attention layers, layer-normalization and position-wise MLPs as follows[2]:

[2]adapted from https://github.com/lucidrains/x-transformers

E. 实现细节与超参数

E.1. 超参数

我们在表12、表13、表14和表15中提供了所有已训练的*LDM*模型的超参数概览。

	CelebA-HQ 256×256	FFHQ 256×256	LSUN-Churches 256×256	LSUN-Bedrooms 256×256
f	4	4	8	4
z-shape	$64 \times 64 \times 3$	$64 \times 64 \times 3$	-	$64 \times 64 \times 3$
$\lvert\mathcal{Z}\rvert$	8192	8192	-	8192
Diffusion steps	1000	1000	1000	1000
Noise Schedule	linear	linear	linear	linear
N_{params}	274M	274M	294M	274M
Channels	224	224	192	224
Depth	2	2	2	2
Channel Multiplier	1,2,3,4	1,2,3,4	1,2,2,4,4	1,2,3,4
Attention resolutions	32, 16, 8	32, 16, 8	32, 16, 8, 4	32, 16, 8
Head Channels	32	32	24	32
Batch Size	48	42	96	48
Iterations*	410k	635k	500k	1.9M
Learning Rate	9.6e-5	8.4e-5	5.e-5	9.6e-5

表12. 用于生成表1中所示数据的无条件*LDMs*超参数。所有模型均在单张NVIDIA A100上训练完成。

	LDM-1	*LDM-2*	*LDM-4*	*LDM-8*	*LDM-16*	*LDM-32*
z-shape	$256 \times 256 \times 3$	$128 \times 128 \times 2$	$64 \times 64 \times 3$	$32 \times 32 \times 4$	$16 \times 16 \times 8$	$88 \times 8 \times 32$
$\lvert\mathcal{Z}\rvert$	-	2048	8192	16384	16384	16384
Diffusion steps	1000	1000	1000	1000	1000	1000
Noise Schedule	linear	linear	linear	linear	linear	linear
Model Size	396M	391M	391M	395M	395M	395M
Channels	192	192	192	256	256	256
Depth	2	2	2	2	2	2
Channel Multiplier	1,1,2,2,4,4	1,2,2,4,4	1,2,3,5	1,2,4	1,2,4	1,2,4
Number of Heads	1	1	1	1	1	1
Batch Size	7	9	40	64	112	112
Iterations	2M	2M	2M	2M	2M	2M
Learning Rate	4.9e-5	6.3e-5	8e-5	6.4e-5	4.5e-5	4.5e-5
Conditioning	CA	CA	CA	CA	CA	CA
CA-resolutions	32, 16, 8	32, 16, 8	32, 16, 8	32, 16, 8	16, 8, 4	8, 4, 2
Embedding Dimension	512	512	512	512	512	512
Transformers Depth	1	1	1	1	1	1

表13. 在ImageNet数据集上训练的条件*LDMs*的超参数，用于第4.1节的分析。所有模型均在单个NVIDIA A100上训练。

E.2. 实现细节

E.2.1 针对条件 *LDMs* 的 τ_θ 实现

在文本到图像和布局到图像（第4.3.1节）合成的实验中，我们将条件生成器 τ_θ 实现为一个无掩码的变换器，该变换器处理输入 y 的标记化版本，并生成输出 $\zeta := \tau_\theta(y)$，其中 $\zeta \in \mathbb{R}^{M \times d_\tau}$。更具体地说，该变换器由 N 个变换器块实现，这些块包含全局自注意力层、层归一化和逐位置MLP，具体结构如下[2]：

[2]adapted from `https://github.com/lucidrains/x-transformers`

	LDM-1	*LDM-2*	*LDM-4*	*LDM-8*	*LDM-16*	*LDM-32*
z-shape	$256 \times 256 \times 3$	$128 \times 128 \times 2$	$64 \times 64 \times 3$	$32 \times 32 \times 4$	$16 \times 16 \times 8$	$88 \times 8 \times 32$
$\lvert\mathcal{Z}\rvert$	-	2048	8192	16384	16384	16384
Diffusion steps	1000	1000	1000	1000	1000	1000
Noise Schedule	linear	linear	linear	linear	linear	linear
Model Size	270M	265M	274M	258M	260M	258M
Channels	192	192	224	256	256	256
Depth	2	2	2	2	2	2
Channel Multiplier	1,1,2,2,4,4	1,2,2,4,4	1,2,3,4	1,2,4	1,2,4	1,2,4
Attention resolutions	32, 16, 8	32, 16, 8	32, 16, 8	32, 16, 8	16, 8, 4	8, 4, 2
Head Channels	32	32	32	32	32	32
Batch Size	9	11	48	96	128	128
Iterations*	500k	500k	500k	500k	500k	500k
Learning Rate	9e-5	1.1e-4	9.6e-5	9.6e-5	1.3e-4	1.3e-4

Table 14. Hyperparameters for the unconditional *LDMs* trained on the CelebA dataset for the analysis in Fig. 7. All models trained on a single NVIDIA A100. *: All models are trained for 500k iterations. If converging earlier, we used the best checkpoint for assessing the provided FID scores.

Task	Text-to-Image	Layout-to-Image		Class-Label-to-Image	Super Resolution	Inpainting	Semantic-Map-to-Image
Dataset	LAION	OpenImages	COCO	ImageNet	ImageNet	Places	Landscapes
f	8	4	8	4	4	4	8
z-shape	$32 \times 32 \times 4$	$64 \times 64 \times 3$	$32 \times 32 \times 4$	$64 \times 64 \times 3$	$64 \times 64 \times 3$	$64 \times 64 \times 3$	$32 \times 32 \times 4$
$\lvert\mathcal{Z}\rvert$	-	8192	16384	8192	8192	8192	16384
Diffusion steps	1000	1000	1000	1000	1000	1000	1000
Noise Schedule	linear	linear	linear	linear	linear	linear	linear
Model Size	1.45B	306M	345M	395M	169M	215M	215M
Channels	320	128	192	192	160	128	128
Depth	2	2	2	2	2	2	2
Channel Multiplier	1,2,4,4	1,2,3,4	1,2,4	1,2,3,5	1,2,2,4	1,4,8	1,4,8
Number of Heads	8	1	1	1	1	1	1
Dropout	-	-	0.1	-	-	-	-
Batch Size	680	24	48	1200	64	128	48
Iterations	390K	4.4M	170K	178K	860K	360K	360K
Learning Rate	1.0e-4	4.8e-5	4.8e-5	1.0e-4	6.4e-5	1.0e-6	4.8e-5
Conditioning	CA	CA	CA	CA	concat	concat	concat
(C)A-resolutions	32, 16, 8	32, 16, 8	32, 16, 8	32, 16, 8	-	-	-
Embedding Dimension	1280	512	512	512	-	-	-
Transformer Depth	1	3	2	1	-	-	-

Table 15. Hyperparameters for the conditional *LDMs* from Sec. 4. All models trained on a single NVIDIA A100 except for the inpainting model which was trained on eight V100.

$$\zeta \leftarrow \text{TokEmb}(y) + \text{PosEmb(y)} \tag{18}$$

$$\text{for } i = 1, \dots, N:$$

$$\zeta_1 \leftarrow \text{LayerNorm}(\zeta) \tag{19}$$

$$\zeta_2 \leftarrow \text{MultiHeadSelfAttention}(\zeta_1) + \zeta \tag{20}$$

$$\zeta_3 \leftarrow \text{LayerNorm}(\zeta_2) \tag{21}$$

$$\zeta \leftarrow \text{MLP}(\zeta_3) + \zeta_2 \tag{22}$$

$$\zeta \leftarrow \text{LayerNorm}(\zeta) \tag{23}$$

$$\tag{24}$$

With ζ available, the conditioning is mapped into the UNet via the cross-attention mechanism as depicted in Fig. 3. We modify the "ablated UNet" [15] architecture and replace the self-attention layer with a shallow (unmasked) transformer consisting of T blocks with alternating layers of (i) self-attention, (ii) a position-wise MLP and (iii) a cross-attention layer;

	LDM-1	*LDM-2*	*LDM-4*	*LDM-8*	*LDM-16*	*LDM-32*
z-shape	$256 \times 256 \times 3$	$128 \times 128 \times 2$	$64 \times 64 \times 3$	$32 \times 32 \times 4$	$16 \times 16 \times 8$	$88 \times 8 \times 32$
$\|\mathcal{Z}\|$	-	2048	8192	16384	16384	16384
Diffusion steps	1000	1000	1000	1000	1000	1000
Noise Schedule	linear	linear	linear	linear	linear	linear
Model Size	270M	265M	274M	258M	260M	258M
Channels	192	192	224	256	256	256
Depth	2	2	2	2	2	2
Channel Multiplier	1,1,2,2,4,4	1,2,2,4,4	1,2,3,4	1,2,4	1,2,4	1,2,4
Attention resolutions	32, 16, 8	32, 16, 8	32, 16, 8	32, 16, 8	16, 8, 4	8, 4, 2
Head Channels	32	32	32	32	32	32
Batch Size	9	11	48	96	128	128
Iterations*	500k	500k	500k	500k	500k	500k
Learning Rate	9e-5	1.1e-4	9.6e-5	9.6e-5	1.3e-4	1.3e-4

表14. 在图7分析中，于CelebA数据集上训练的无条件*LDMs*超参数。所有模型均在单张NVIDIA A100上训练。*：所有模型均训练50万次迭代。若提前收敛，则使用最佳检查点评估所提供的FID分数。

Task	Text-to-Image	Layout-to-Image		Class-Label-to-Image	Super Resolution	Inpainting	Semantic-Map-to-Image
Dataset	LAION	OpenImages	COCO	ImageNet	ImageNet	Places	Landscapes
f	8	4	8	4	4	4	8
z-shape	$32 \times 32 \times 4$	$64 \times 64 \times 3$	$32 \times 32 \times 4$	$64 \times 64 \times 3$	$64 \times 64 \times 3$	$64 \times 64 \times 3$	$32 \times 32 \times 4$
$\|\mathcal{Z}\|$	-	8192	16384	8192	8192	8192	16384
Diffusion steps	1000	1000	1000	1000	1000	1000	1000
Noise Schedule	linear	linear	linear	linear	linear	linear	linear
Model Size	1.45B	306M	345M	395M	169M	215M	215M
Channels	320	128	192	192	160	128	128
Depth	2	2	2	2	2	2	2
Channel Multiplier	1,2,4,4	1,2,3,4	1,2,4	1,2,3,5	1,2,2,4	1,4,8	1,4,8
Number of Heads	8	1	1	1	1	1	1
Dropout	-	-	0.1	-	-	-	-
Batch Size	680	24	48	1200	64	128	48
Iterations	390K	4.4M	170K	178K	860K	360K	360K
Learning Rate	1.0e-4	4.8e-5	4.8e-5	1.0e-4	6.4e-5	1.0e-6	4.8e-5
Conditioning	CA	CA	CA	CA	concat	concat	concat
(C)A-resolutions	32, 16, 8	32, 16, 8	32, 16, 8	32, 16, 8	-	-	-
Embedding Dimension	1280	512	512	512	-	-	-
Transformer Depth	1	3	2	1	-	-	-

表15. 第4节中条件*LDMs*的超参数。除修复模型在八块V100上训练外，其余模型均在单块NVIDIA A100上完成训练。

$$\zeta \leftarrow \text{TokEmb}(y) + \text{PosEmb(y)} \tag{18}$$

$$\text{for } i = 1, \dots, N:$$

$$\zeta_1 \leftarrow \text{LayerNorm}(\zeta) \tag{19}$$

$$\zeta_2 \leftarrow \text{MultiHeadSelfAttention}(\zeta_1) + \zeta \tag{20}$$

$$\zeta_3 \leftarrow \text{LayerNorm}(\zeta_2) \tag{21}$$

$$\zeta \leftarrow \text{MLP}(\zeta_3) + \zeta_2 \tag{22}$$

$$\zeta \leftarrow \text{LayerNorm}(\zeta) \tag{23}$$

$$\tag{24}$$

在ζ可用的情况下，条件信息通过交叉注意力机制映射到UNet中，如图3所示。我们修改了“简化版UNet”[15]架构，将自注意力层替换为一个浅层（非掩码）Transformer，该结构包含T个交替堆叠的模块，每个模块依次包含：(i)自注意力层、(ii)逐位置MLP以及(iii)交叉注意力层；

see Tab. 16. Note that without (ii) and (iii), this architecture is equivalent to the "ablated UNet".

While it would be possible to increase the representational power of τ_θ by additionally conditioning on the time step t, we do not pursue this choice as it reduces the speed of inference. We leave a more detailed analysis of this modification to future work.

For the text-to-image model, we rely on a publicly available[3] tokenizer [99]. The layout-to-image model discretizes the spatial locations of the bounding boxes and encodes each box as a (l, b, c)-tuple, where l denotes the (discrete) top-left and b the bottom-right position. Class information is contained in c.
See Tab. 17 for the hyperparameters of τ_θ and Tab. 13 for those of the UNet for both of the above tasks.

Note that the class-conditional model as described in Sec. 4.1 is also implemented via cross-attention, where τ_θ is a single learnable embedding layer with a dimensionality of 512, mapping classes y to $\zeta \in \mathbb{R}^{1\times 512}$.

input	$\mathbb{R}^{h\times w\times c}$
LayerNorm	$\mathbb{R}^{h\times w\times c}$
Conv1x1	$\mathbb{R}^{h\times w\times d\cdot n_h}$
Reshape	$\mathbb{R}^{h\cdot w\times d\cdot n_h}$
$\times T$ { SelfAttention	$\mathbb{R}^{h\cdot w\times d\cdot n_h}$
$\times T$ { MLP	$\mathbb{R}^{h\cdot w\times d\cdot n_h}$
$\times T$ { CrossAttention	$\mathbb{R}^{h\cdot w\times d\cdot n_h}$
Reshape	$\mathbb{R}^{h\times w\times d\cdot n_h}$
Conv1x1	$\mathbb{R}^{h\times w\times c}$

Table 16. Architecture of a transformer block as described in Sec. E.2.1, replacing the self-attention layer of the standard "ablated UNet" architecture [15]. Here, n_h denotes the number of attention heads and d the dimensionality per head.

	Text-to-Image	**Layout-to-Image**
seq-length	77	92
depth N	32	16
dim	1280	512

Table 17. Hyperparameters for the experiments with transformer encoders in Sec. 4.3.

E.2.2 Inpainting

For our experiments on image-inpainting in Sec. 4.5, we used the code of [88] to generate synthetic masks. We use a fixed set of 2k validation and 30k testing samples from Places [108]. During training, we use random crops of size 256×256 and evaluate on crops of size 512×512. This follows the training and testing protocol in [88] and reproduces their reported metrics (see † in Tab. 7). We include additional qualitative results of *LDM-4, w/ attn* in Fig. 21 and of *LDM-4, w/o attn, big, w/ ft* in Fig. 22.

E.3. Evaluation Details

This section provides additional details on evaluation for the experiments shown in Sec. 4.

E.3.1 Quantitative Results in Unconditional and Class-Conditional Image Synthesis

We follow common practice and estimate the statistics for calculating the FID-, Precision- and Recall-scores [29,50] shown in Tab. 1 and 10 based on 50k samples from our models and the entire training set of each of the shown datasets. For calculating FID scores we use the `torch-fidelity` package [60]. However, since different data processing pipelines might lead to different results [64], we also evaluate our models with the script provided by Dhariwal and Nichol [15]. We find that results

[3] https://huggingface.co/transformers/model_doc/bert.html#berttokenizerfast

参见表16。注意，如果没有（ii）和（iii），该架构等同于“消融的UNet”。

虽然可以通过额外引入时间步t的条件来增强τ_θ的表征能力，但我们未采用这一方案，因为它会降低推理速度。我们将对此修改的更详细分析留待未来工作。

对于文本到图像模型，我们采用公开可用的[3]分词器[99]。布局到图像模型将边界框的空间位置离散化，并将每个框编码为一个(l, b, c)元组，其中l表示（离散化的）左上角位置，b表示右下角位置。类别信息包含在c中。

请参见表 17 了解 τ_θ 的超参数，并参见表 13 了解上述两个任务中 UNet 的超参数。

请注意，第4.1节中描述的类条件模型同样通过交叉注意力机制实现，其中τ_θ是一个维度为512的可学习嵌入层，将类别y映射到$\zeta \in \mathbb{R}^{1\times 512}$。

input	$\mathbb{R}^{h\times w\times c}$
LayerNorm	$\mathbb{R}^{h\times w\times c}$
Conv1x1	$\mathbb{R}^{h\times w\times d\cdot n_h}$
Reshape	$\mathbb{R}^{h\cdot w\times d\cdot n_h}$
$\times T$ { SelfAttention	$\mathbb{R}^{h\cdot w\times d\cdot n_h}$
$\times T$ { MLP	$\mathbb{R}^{h\cdot w\times d\cdot n_h}$
$\times T$ { CrossAttention	$\mathbb{R}^{h\cdot w\times d\cdot n_h}$
Reshape	$\mathbb{R}^{h\times w\times d\cdot n_h}$
Conv1x1	$\mathbb{R}^{h\times w\times c}$

表16. 如第E.2.1节所述，替换标准“消融UNet”架构[15]中自注意力层的Transformer块架构。此处，n_h表示注意力头数量，d表示每个头的维度。

	Text-to-Image	**Layout-to-Image**
seq-length	77	92
depth N	32	16
dim	1280	512

表 17. 第 4.3 节中 Transformer 编码器实验的超参数。

E.2.2 图像修复

在4.5节关于图像修复的实验中，我们采用[88]的代码生成合成掩码。我们从Places数据集[108]中选取固定的2k张验证样本和30k张测试样本。训练时使用256×256的随机裁剪图像，评估时则采用512×512的裁剪尺寸。该流程遵循[88]的训练测试协议，并复现了其报告指标（见表7中的†）。我们在图21中补充了*LDM-4, w/ attn*的定性结果，在图22中补充了*LDM-4, w/o attn, big,w/ ft*的定性结果。

E.3. 评估详情

本节提供了关于第4节所示实验评估的更多细节。

E.3.1 无条件与类别条件图像合成的定量结果

我们遵循常规做法，基于从我们模型中抽取的5万个样本以及每个所示数据集的完整训练集，估算用于计算表1和表10中FID、精确度和召回率分数[29,50]的统计量。在计算FID分数时，我们使用torch-fidelity包[60]。然而，由于不同的数据处理流程可能导致不同结果[64]，我们还使用Dhariwal和Nichol[15]提供的脚本评估了我们的模型。我们发现结果

[3] https://huggingface.co/transformers/model_doc/bert.html#berttokenizerfast

mainly coincide, except for the ImageNet and LSUN-Bedrooms datasets, where we notice slightly varying scores of 7.76 (`torch-fidelity`) vs. 7.77 (Nichol and Dhariwal) and 2.95 vs 3.0. For the future we emphasize the importance of a unified procedure for sample quality assessment. Precision and Recall are also computed by using the script provided by Nichol and Dhariwal.

E.3.2 Text-to-Image Synthesis

Following the evaluation protocol of [66] we compute FID and Inception Score for the Text-to-Image models from Tab. 2 by comparing generated samples with 30000 samples from the validation set of the MS-COCO dataset [51]. FID and Inception Scores are computed with `torch-fidelity`.

E.3.3 Layout-to-Image Synthesis

For assessing the sample quality of our Layout-to-Image models from Tab. 9 on the COCO dataset, we follow common practice [37, 87, 89] and compute FID scores the 2048 unaugmented examples of the COCO Segmentation Challenge split. To obtain better comparability, we use the exact same samples as in [37]. For the OpenImages dataset we similarly follow their protocol and use 2048 center-cropped test images from the validation set.

E.3.4 Super Resolution

We evaluate the super-resolution models on ImageNet following the pipeline suggested in [72], *i.e*. images with a shorter size less than 256 px are removed (both for training and evaluation). On ImageNet, the low-resolution images are produced using bicubic interpolation with anti-aliasing. FIDs are evaluated using `torch-fidelity` [60], and we produce samples on the validation split. For FID scores, we additionally compare to reference features computed on the train split, see Tab. 5 and Tab. 11.

E.3.5 Efficiency Analysis

For efficiency reasons we compute the sample quality metrics plotted in Fig. 6, 17 and 7 based on 5k samples. Therefore, the results might vary from those shown in Tab. 1 and 10. All models have a comparable number of parameters as provided in Tab. 13 and 14. We maximize the learning rates of the individual models such that they still train stably. Therefore, the learning rates slightly vary between different runs *cf*. Tab. 13 and 14.

E.3.6 User Study

For the results of the user study presented in Tab. 4 we followed the protocoll of [72] and and use the 2-alternative force-choice paradigm to assess human preference scores for two distinct tasks. In Task-1 subjects were shown a low resolution/masked image between the corresponding ground truth high resolution/unmasked version and a synthesized image, which was generated by using the middle image as conditioning. For SuperResolution subjects were asked: *'Which of the two images is a better high quality version of the low resolution image in the middle?'*. For Inpainting we asked *'Which of the two images contains more realistic inpainted regions of the image in the middle?'*. In Task-2, humans were similarly shown the low-res/masked version and asked for preference between two corresponding images generated by the two competing methods. As in [72] humans viewed the images for 3 seconds before responding.

主要重合，除了ImageNet和LSUN-Bedrooms数据集，我们注意到略有不同的分数：7.76（torch-fidelity）对比7.77（Nichol and Dhariwal），以及2.95对比3.0。未来我们强调样本质量评估统一流程的重要性。精确率和召回率同样使用Nichol和Dhariwal提供的脚本进行计算。

E.3.2 文本到图像合成

遵循[66]的评估协议，我们通过将生成样本与MS-COCO数据集[51]验证集中的30000个样本进行比较，计算了表2中文本到图像模型的FID和初始分数。FID和初始分数使用torch-fidelity计算。

E.3.3 布局到图像合成

为了评估我们在COCO数据集上基于表格9的布局到图像模型的样本质量，我们遵循常见做法[37, 87, 89]，使用COCO分割挑战赛划分中2048个未经增强的示例计算FID分数。为了获得更好的可比性，我们采用与[37]完全相同的样本。对于OpenImages数据集，我们同样遵循其协议，使用验证集中的2048张中心裁剪测试图像。

E.3.4 超分辨率

我们按照[72]中建议的流程在ImageNet上评估超分辨率模型，*i.e*。对于训练和评估，我们移除了短边小于256像素的图像。在ImageNet上，低分辨率图像是通过带抗锯齿的双三次插值生成的。FID分数使用torch-fidelity [60]进行评估，并在验证集上生成样本。对于FID分数，我们还额外与在训练集上计算的参考特征进行了比较，详见表5和表11。

E.3.5 效率分析

出于效率考虑，我们基于5千个样本计算了图6、17和7中绘制的样本质量指标。因此，结果可能与表1和表10中所示存在差异。所有模型的参数量级均与表13和表14中提供的数据相当。我们将各模型的初始学习率调整至仍能保持稳定训练的最大值，因此不同实验*cf*的学习率会略有浮动。表13和表14。

E.3.6 用户研究

对于表4中展示的用户研究结果，我们遵循[72]的协议，采用二选一强制选择范式来评估人类对两项独立任务的偏好得分。在任务1中，受试者会看到一张低分辨率/遮挡图像，并同时展示对应的高分辨率/未遮挡真实版本图像与一张合成图像——后者是以中间图像为条件生成的。对于超分辨率任务，我们询问受试者：
'Which of the two images is a better high quality version of the low resolution image in the middle?'；对于图像修复任务，我们则询问*'Which of the two images contains more realistic inpainted regions of the image in the middle?'*。在任务2中，受试者同样会看到低分辨率/遮挡版本图像，并被要求对两种竞争方法生成的对应图像进行偏好选择。如[72]所述，受试者在回答前有3秒时间观察图像。

F. Computational Requirements

Method	Generator Compute	Classifier Compute	Overall Compute	Inference Throughput*	N_{params}	FID↓	IS↑	Precision↑	Recall↑
LSUN Churches 256^2									
StyleGAN2 [42]†	64	-	64	-	59M	3.86	-	-	-
LDM-8 (ours, 100 steps, 410K)	18	-	18	6.80	256M	4.02	-	0.64	0.52
LSUN Bedrooms 256^2									
ADM [15]† (1000 steps)	232	-	232	0.03	552M	1.9	-	0.66	0.51
LDM-4 (ours, 200 steps, 1.9M)	60	-	55	1.07	274M	2.95	-	0.66	0.48
CelebA-HQ 256^2									
LDM-4 (ours, 500 steps, 410K)	14.4	-	14.4	0.43	274M	5.11	-	0.72	0.49
FFHQ 256^2									
StyleGAN2 [42]	32.13‡	-	32.13†	-	59M	3.8	-	-	-
LDM-4 (ours, 200 steps, 635K)	26	-	26	1.07	274M	4.98	-	0.73	0.50
ImageNet 256^2									
VQGAN-f-4 (ours, first stage)	29	-	29	-	55M	0.58††	-	-	-
VQGAN-f-8 (ours, first stage)	66	-	66	-	68M	1.14††	-	-	-
BigGAN-deep [3]†	128-256		128-256	-	340M	6.95	$203.6_{\pm 2.6}$	0.87	0.28
ADM [15] (250 steps) †	916	-	916	0.12	554M	10.94	100.98	0.69	0.63
ADM-G [15] (25 steps) †	916	46	962	0.7	608M	5.58	-	0.81	0.49
ADM-G [15] (250 steps)†	916	46	962	0.07	608M	4.59	186.7	0.82	0.52
ADM-G,ADM-U [15] (250 steps)†	329	30	349	n/a	n/a	3.85	221.72	0.84	0.53
LDM-8-G (ours, 100, 2.9M)	79	12	91	1.93	506M	8.11	$190.4_{\pm 2.6}$	0.83	0.36
LDM-8 (ours, 200 ddim steps 2.9M, batch size 64)	79	-	79	1.9	395M	17.41	72.92	0.65	0.62
LDM-4 (ours, 250 ddim steps 178K, batch size 1200)	271	-	271	0.7	400M	10.56	$103.49_{\pm 1.24}$	0.71	0.62
LDM-4-G (ours, 250 ddim steps 178K, batch size 1200, classifier-free guidance [32] scale 1.25)	271	-	271	0.4	400M	3.95	$178.22_{\pm 2.43}$	0.81	0.55
LDM-4-G (ours, 250 ddim steps 178K, batch size 1200, classifier-free guidance [32] scale 1.5)	271	-	271	0.4	400M	3.60	$247.67_{\pm 5.59}$	0.87	0.48

Table 18. Comparing compute requirements during training and inference throughput with state-of-the-art generative models. Compute during training in V100-days, numbers of competing methods taken from [15] unless stated differently;*: Throughput measured in samples/sec on a single NVIDIA A100;†: Numbers taken from [15] ;‡: Assumed to be trained on 25M train examples; ††: R-FID vs. ImageNet validation set

In Tab 18 we provide a more detailed analysis on our used compute ressources and compare our best performing models on the CelebA-HQ, FFHQ, LSUN and ImageNet datasets with the recent state of the art models by using their provided numbers, *cf*. [15]. As they report their used compute in V100 days and we train all our models on a single NVIDIA A100 GPU, we convert the A100 days to V100 days by assuming a $\times 2.2$ speedup of A100 vs V100 [74][4]. To assess sample quality, we additionally report FID scores on the reported datasets. We closely reach the performance of state of the art methods as StyleGAN2 [42] and ADM [15] while significantly reducing the required compute resources.

[4]This factor corresponds to the speedup of the A100 over the V100 for a U-Net, as defined in Fig. 1 in [74]

F. 计算要求

Method	Generator Compute	Classifier Compute	Overall Compute	Inference Throughput*	N_{params}	FID↓	IS↑	Precision↑	Recall↑
LSUN Churches 256^2									
StyleGAN2 [42]†	64	-	64	-	59M	3.86	-	-	-
LDM-8 (ours, 100 steps, 410K)	18	-	18	6.80	256M	4.02	-	0.64	0.52
LSUN Bedrooms 256^2									
ADM [15]† (1000 steps)	232	-	232	0.03	552M	1.9	-	0.66	0.51
LDM-4 (ours, 200 steps, 1.9M)	60	-	55	1.07	274M	2.95	-	0.66	0.48
CelebA-HQ 256^2									
LDM-4 (ours, 500 steps, 410K)	14.4	-	14.4	0.43	274M	5.11	-	0.72	0.49
FFHQ 256^2									
StyleGAN2 [42]	32.13‡	-	32.13†	-	59M	3.8	-	-	-
LDM-4 (ours, 200 steps, 635K)	26	-	26	1.07	274M	4.98	-	0.73	0.50
ImageNet 256^2									
VQGAN-f-4 (ours, first stage)	29	-	29	-	55M	0.58††	-	-	-
VQGAN-f-8 (ours, first stage)	66	-	66	-	68M	1.14††	-	-	-
BigGAN-deep [3]†	128-256		128-256	-	340M	6.95	$203.6_{\pm 2.6}$	0.87	0.28
ADM [15] (250 steps) †	916	-	916	0.12	554M	10.94	100.98	0.69	0.63
ADM-G [15] (25 steps) †	916	46	962	0.7	608M	5.58	-	0.81	0.49
ADM-G [15] (250 steps)†	916	46	962	0.07	608M	4.59	186.7	0.82	0.52
ADM-G,ADM-U [15] (250 steps)†	329	30	349	n/a	n/a	3.85	221.72	0.84	0.53
LDM-8-G (ours, 100, 2.9M)	79	12	91	1.93	506M	8.11	$190.4_{\pm 2.6}$	0.83	0.36
LDM-8 (ours, 200 ddim steps 2.9M, batch size 64)	79	-	79	1.9	395M	17.41	72.92	0.65	0.62
LDM-4 (ours, 250 ddim steps 178K, batch size 1200)	271	-	271	0.7	400M	10.56	$103.49_{\pm 1.24}$	0.71	0.62
LDM-4-G (ours, 250 ddim steps 178K, batch size 1200, classifier-free guidance [32] scale 1.25)	271	-	271	0.4	400M	3.95	$178.22_{\pm 2.43}$	0.81	0.55
LDM-4-G (ours, 250 ddim steps 178K, batch size 1200, classifier-free guidance [32] scale 1.5)	271	-	271	0.4	400M	3.60	$247.67_{\pm 5.59}$	0.87	0.48

表18. 比较训练期间的计算需求与最先进生成模型的推理吞吐量。训练计算量以V100-天为单位，竞争方法的数据取自[15]（除非另有说明）；*：吞吐量在单个NVIDIA A100上以样本/秒为单位测量；†：数据取自[15]；‡：假设在2500万训练样本上训练；††：R-FID与ImageNet验证集的对比结果

在表18中，我们对所使用的计算资源进行了更详细的分析，并通过使用现有最新模型提供的数值*cf*，将我们在CelebA-HQ、FFHQ、LSUN和ImageNet数据集上表现最佳的模型与这些模型进行了比较[15]。由于他们报告的计算资源使用量以V100天为单位，而我们在单个NVIDIA A100 GPU上训练所有模型，因此我们通过假设A100相比V100有$\times 2.2$倍的加速比[74][4]，将A100天数转换为V100天数。为了评估样本质量，我们还报告了相关数据集的FID分数。在显著减少所需计算资源的同时，我们几乎达到了StyleGAN2 [42]和ADM [15]等最先进方法的性能水平。

[4]This factor corresponds to the speedup of the A100 over the V100 for a U-Net, as defined in Fig. 1 in [74]

G. Details on Autoencoder Models

We train all our autoencoder models in an adversarial manner following [23], such that a patch-based discriminator D_ψ is optimized to differentiate original images from reconstructions $\mathcal{D}(\mathcal{E}(x))$. To avoid arbitrarily scaled latent spaces, we regularize the latent z to be zero centered and obtain small variance by introducing an regularizing loss term L_{reg}.
We investigate two different regularization methods: (i) a low-weighted Kullback-Leibler-term between $q_\mathcal{E}(z|x) = \mathcal{N}(z; \mathcal{E}_\mu, \mathcal{E}_{\sigma^2})$ and a standard normal distribution $\mathcal{N}(z; 0, 1)$ as in a standard variational autoencoder [46, 69], and, (ii) regularizing the latent space with a vector quantization layer by learning a codebook of $|\mathcal{Z}|$ different exemplars [96].
To obtain high-fidelity reconstructions we only use a very small regularization for both scenarios, *i.e.* we either weight the $\mathbb{KL}$ term by a factor $\sim 10^{-6}$ or choose a high codebook dimensionality $|\mathcal{Z}|$.

The full objective to train the autoencoding model $(\mathcal{E}, \mathcal{D})$ reads:

$$L_{\text{Autoencoder}} = \min_{\mathcal{E},\mathcal{D}} \max_{\psi} \Big(L_{rec}(x, \mathcal{D}(\mathcal{E}(x))) - L_{adv}(\mathcal{D}(\mathcal{E}(x))) + \log D_\psi(x) + L_{reg}(x; \mathcal{E}, \mathcal{D}) \Big) \tag{25}$$

DM Training in Latent Space Note that for training diffusion models on the learned latent space, we again distinguish two cases when learning $p(z)$ or $p(z|y)$ (Sec. 4.3): (i) For a KL-regularized latent space, we sample $z = \mathcal{E}_\mu(x) + \mathcal{E}_\sigma(x) \cdot \varepsilon =: \mathcal{E}(x)$, where $\varepsilon \sim \mathcal{N}(0, 1)$. When rescaling the latent, we estimate the component-wise variance

$$\hat{\sigma}^2 = \frac{1}{bchw} \sum_{b,c,h,w} (z^{b,c,h,w} - \hat{\mu})^2$$

from the first batch in the data, where $\hat{\mu} = \frac{1}{bchw} \sum_{b,c,h,w} z^{b,c,h,w}$. The output of $\mathcal{E}$ is scaled such that the rescaled latent has unit standard deviation, *i.e.* $z \leftarrow \frac{z}{\hat{\sigma}} = \frac{\mathcal{E}(x)}{\hat{\sigma}}$. (ii) For a VQ-regularized latent space, we extract z *before* the quantization layer and absorb the quantization operation into the decoder, *i.e.* it can be interpreted as the first layer of $\mathcal{D}$.

H. Additional Qualitative Results

Finally, we provide additional qualitative results for our landscapes model (Fig. 12, 23, 24 and 25), our class-conditional ImageNet model (Fig. 26 - 27) and our unconditional models for the CelebA-HQ, FFHQ and LSUN datasets (Fig. 28 - 31). Similar as for the inpainting model in Sec. 4.5 we also fine-tuned the semantic landscapes model from Sec. 4.3.2 directly on 512^2 images and depict qualitative results in Fig. 12 and Fig. 23. For our those models trained on comparably small datasets, we additionally show nearest neighbors in VGG [79] feature space for samples from our models in Fig. 32 - 34.

G. 自编码器模型详情

我们按照[23]的方式以对抗方式训练所有自编码器模型，使得基于块的判别器D_ψ被优化以区分原始图像和重建图像$\mathcal{D}(\mathcal{E}(x))$。为避免潜在空间任意缩放，我们通过引入正则化损失项L_{reg}来约束潜在变量z以零为中心并保持较小方差。我们研究了两种不同的正则化方法：(i) 采用低权重的Kullback-Leibler项，计算$q_\mathcal{E}(z|x) = \mathcal{N}(z; \mathcal{E}_\mu, \mathcal{E}_{\sigma^2})$与标准正态分布$\mathcal{N}(z$之间的差异，即0,1)，如标准变分自编码器[46, 69]所示；以及(ii) 通过学习包含$|\mathcal{Z}|$个不同样本的码本[96]，利用向量量化层对潜在空间进行正则化。为获得高保真重建，我们在两种场景下均仅使用极弱的正则化强度*i.e*：要么将$\mathbb{KL}$项的权重设为因子$\sim 10^{-6}$，要么选择较高的码本维度$|\mathcal{Z}|$。

训练自编码模型（{v*}）的完整目标函数如下：

$$L_{\text{Autoencoder}} = \min_{\mathcal{E},\mathcal{D}} \max_{\psi} \Big(L_{rec}(x, \mathcal{D}(\mathcal{E}(x))) - L_{adv}(\mathcal{D}(\mathcal{E}(x))) + \log D_\psi(x) + L_{reg}(x; \mathcal{E}, \mathcal{D}) \Big) \tag{25}$$

潜在空间中的扩散模型训练 请注意，对于在已学习的潜在空间上训练扩散模型，在学习$p(z)$或$p(z|y)$ (第4.3节)时，我们再次区分两种情况：（i）对于KL正则化的潜在空间，我们采样$z = \mathcal{E}_\mu(x) + \mathcal{E}_\sigma(x) \cdot \varepsilon =: \mathcal{E}(x)$，其中$\varepsilon \sim \mathcal{N}(0, 1)$。在重新缩放潜在表示时，我们估计分量方差

$$\hat{\sigma}^2 = \frac{1}{bchw} \sum_{b,c,h,w} (z^{b,c,h,w} - \hat{\mu})^2$$

从数据的第一批开始，其中$\hat{\mu} = \frac{1}{bchw} \sum_{b,c,h,w} z^{b,c,h,w}$。$\mathcal{E}$的输出经过缩放，使得重新缩放的潜变量具有单位标准差，*i.e*。$z \leftarrow \frac{z}{\hat{\sigma}} = \frac{\mathcal{E}(x)}{\hat{\sigma}}$。（ii）对于VQ正则化的潜空间，我们提取$z$ *before*量化层并将量化操作吸收到解码器中，*i.e*。这可以解释为$\mathcal{D}$的第一层。

H. 其他定性结果

最后，我们为我们的景观模型（图12、23、24和25）、我们的类别条件ImageNet模型（图26-27）以及我们在CelebA-HQ、FFHQ和LSUN数据集上的无条件模型（图28-31）提供了额外的定性结果。与第4.5节中的修复模型类似，我们也直接在第4.3.2节的语义景观模型上对512^2图像进行了微调，并在图12和图23中展示了定性结果。对于那些在相对较小数据集上训练的模型，我们还在图32-34中展示了模型样本在VGG[79]特征空间中的最近邻。

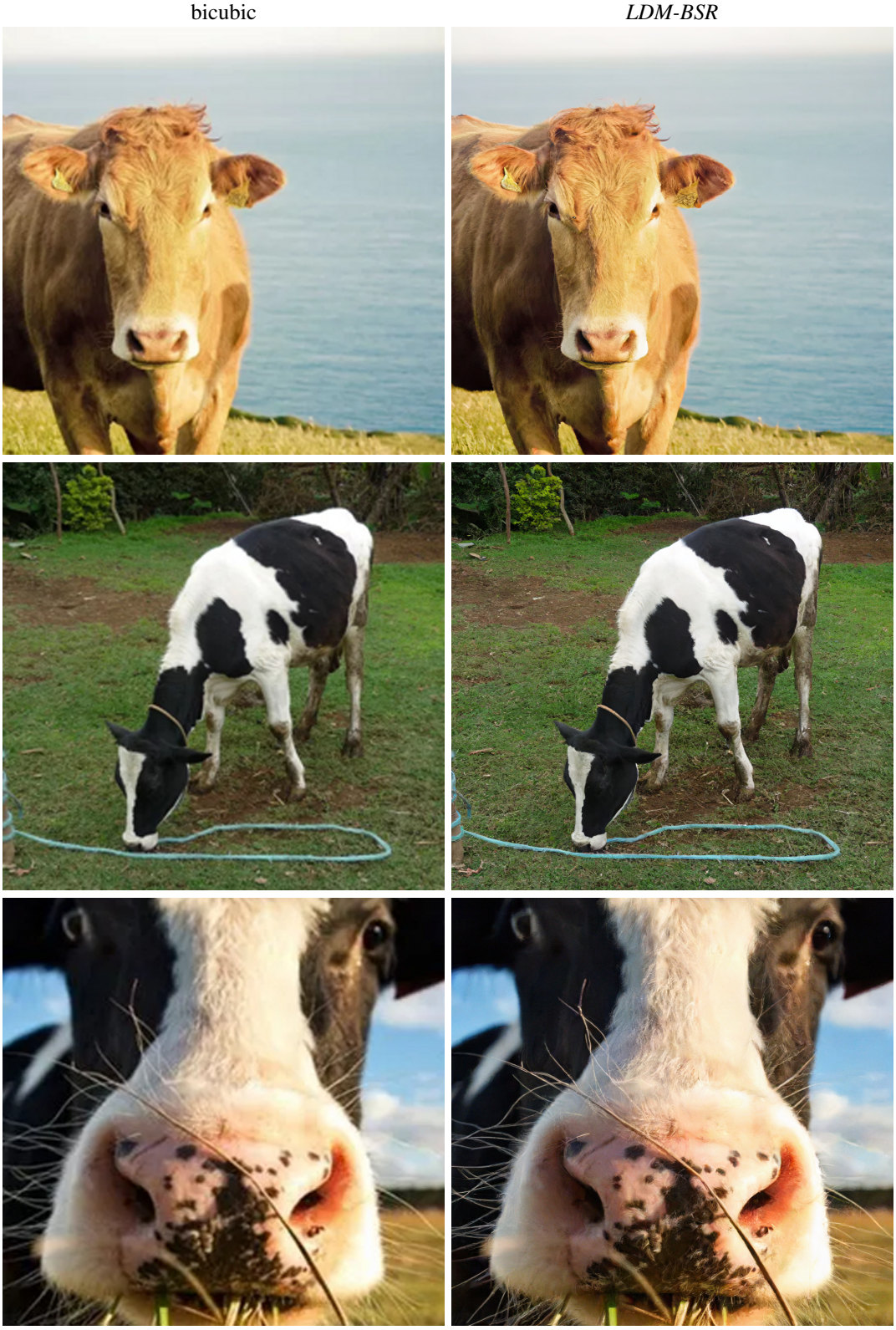

Figure 19. *LDM-BSR* generalizes to arbitrary inputs and can be used as a general-purpose upsampler, upscaling samples from the LSUN-Cows dataset to 1024^2 resolution.

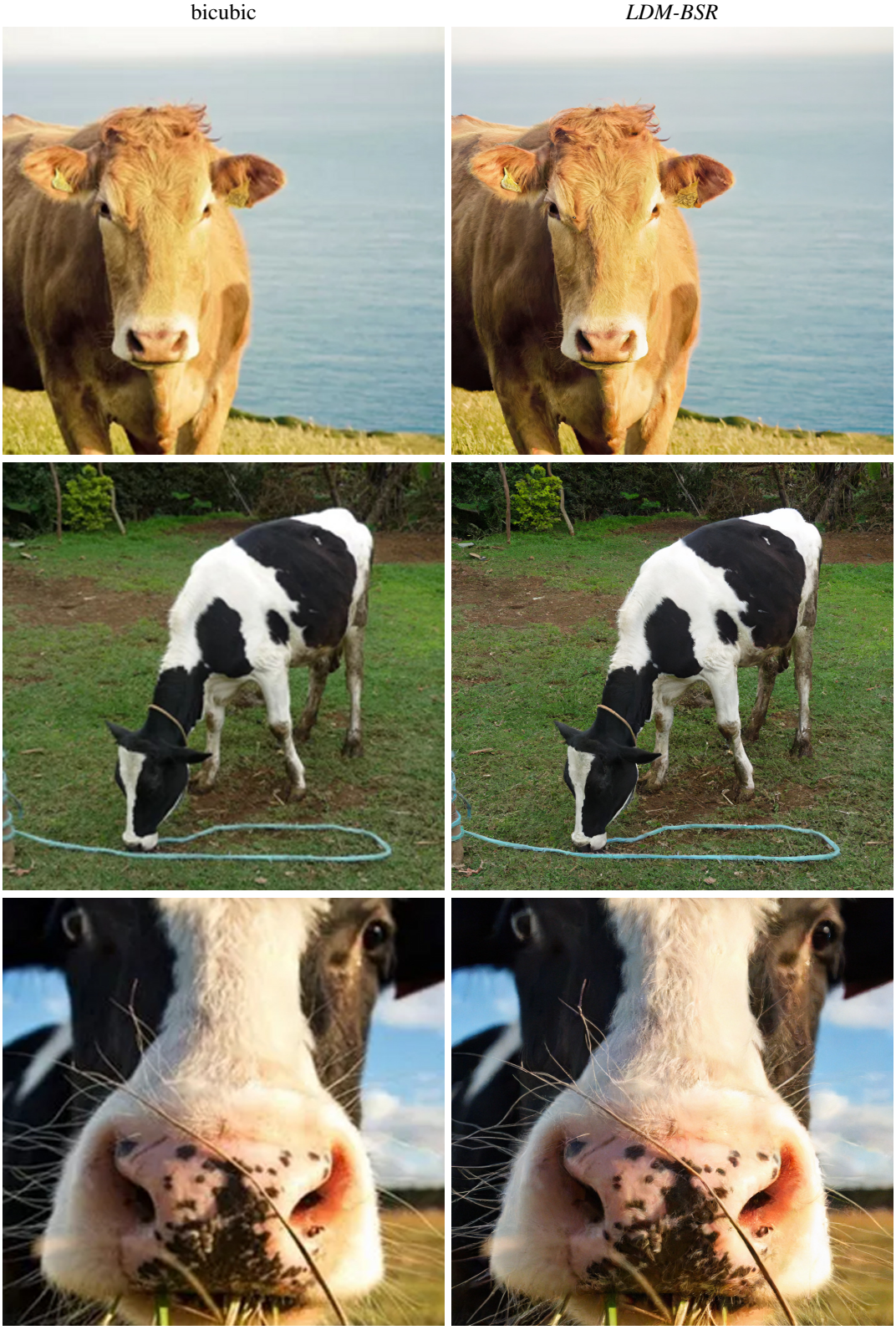

图19。*LDM-BSR* 可泛化至任意输入，并能作为通用上采样器使用，将LSUN-Cows数据集的样本提升至1024^2分辨率。

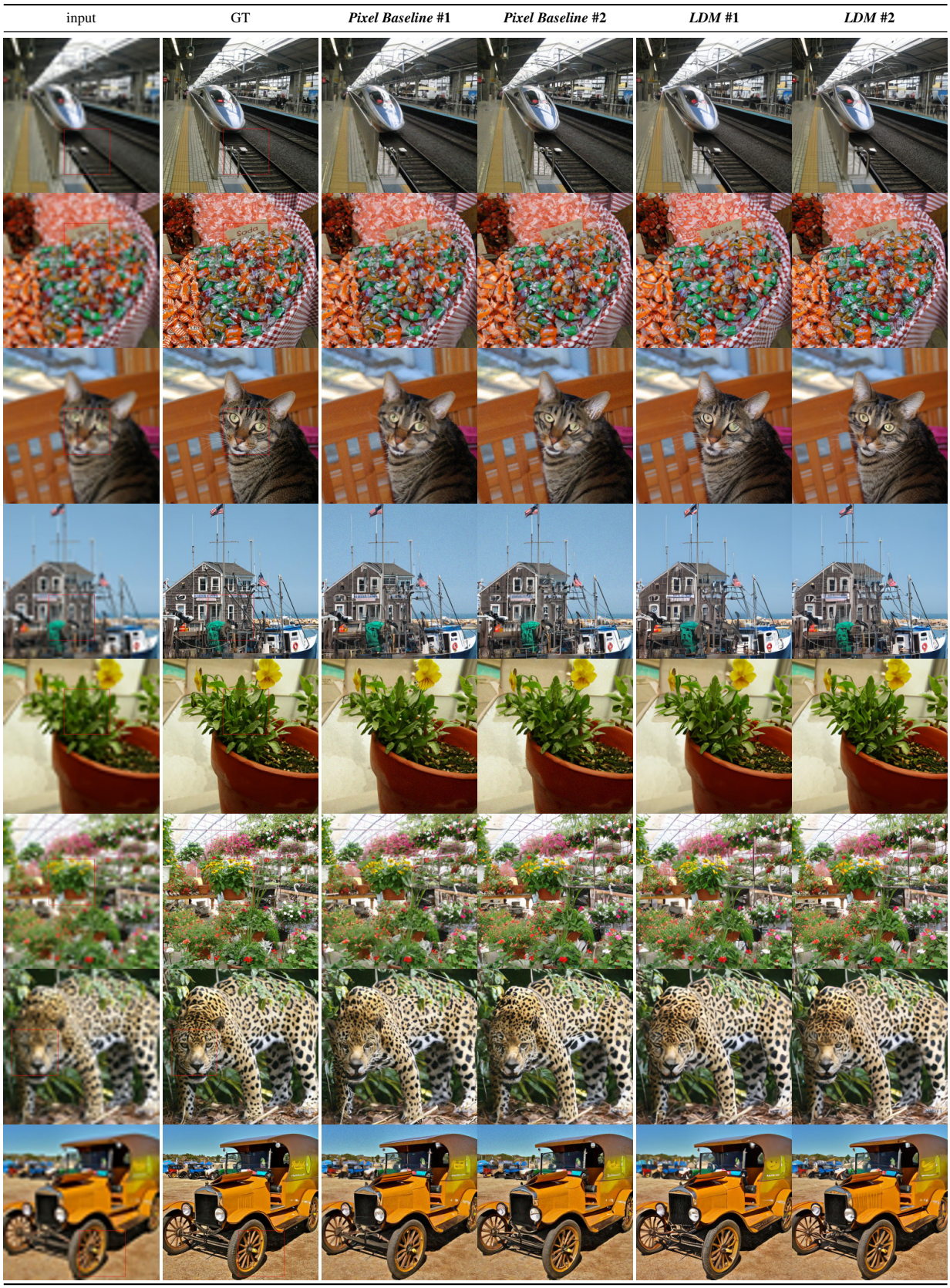

Figure 20. Qualitative superresolution comparison of two random samples between LDM-SR and baseline-diffusionmodel in Pixelspace. Evaluated on imagenet validation-set after same amount of training steps.

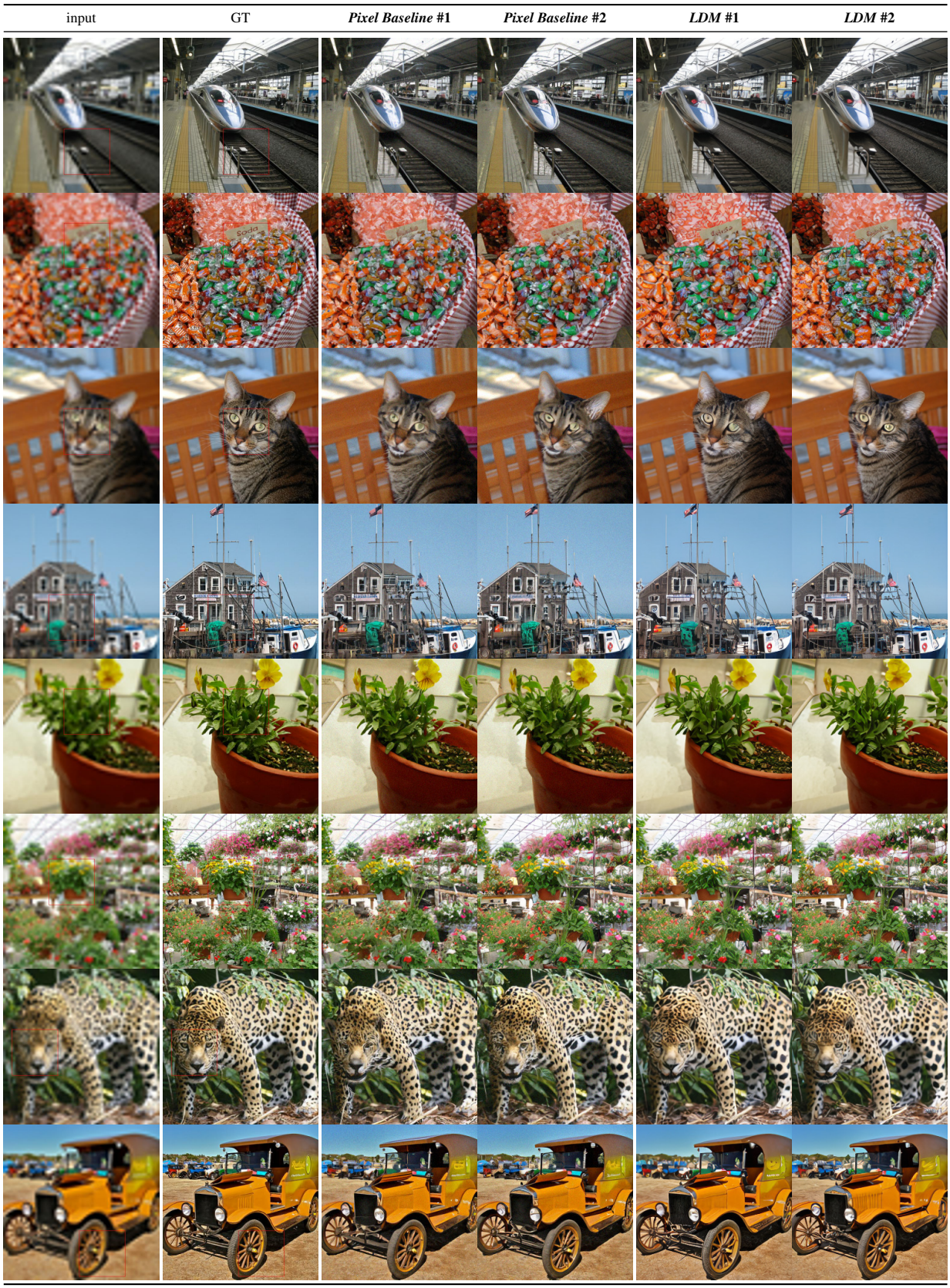

图20. LDM-SR与基线扩散模型在像素空间中对两个随机样本的定性超分辨率比较。在相同训练步数后，基于ImageNet验证集进行评估。

Figure 21. Qualitative results on image inpainting. In contrast to [88], our generative approach enables generation of multiple diverse samples for a given input.

图21. 图像修复的定性结果。与[88]相比，我们的生成方法能够为给定输入生成多个多样化的样本。

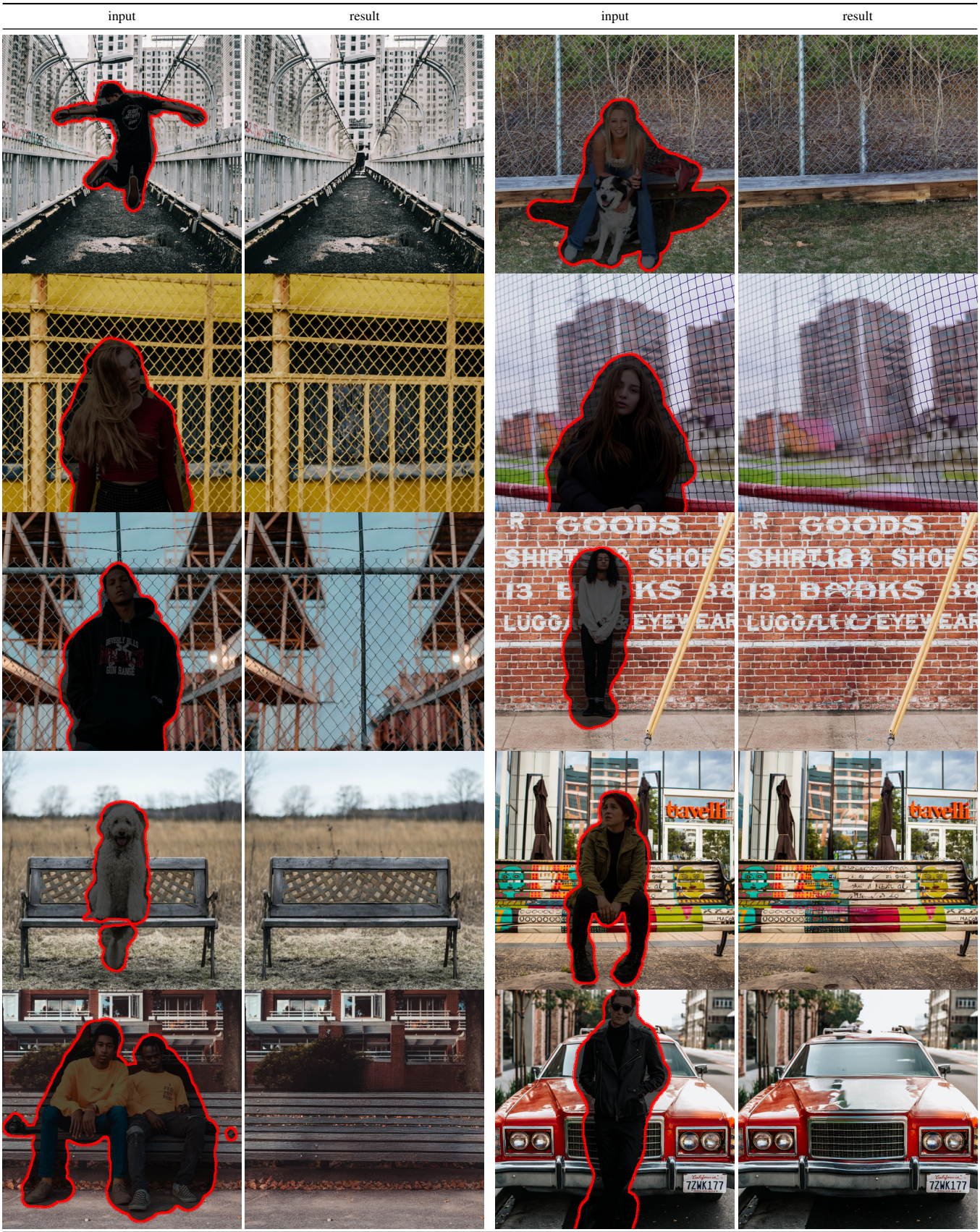

Figure 22. More qualitative results on object removal as in Fig. 11.

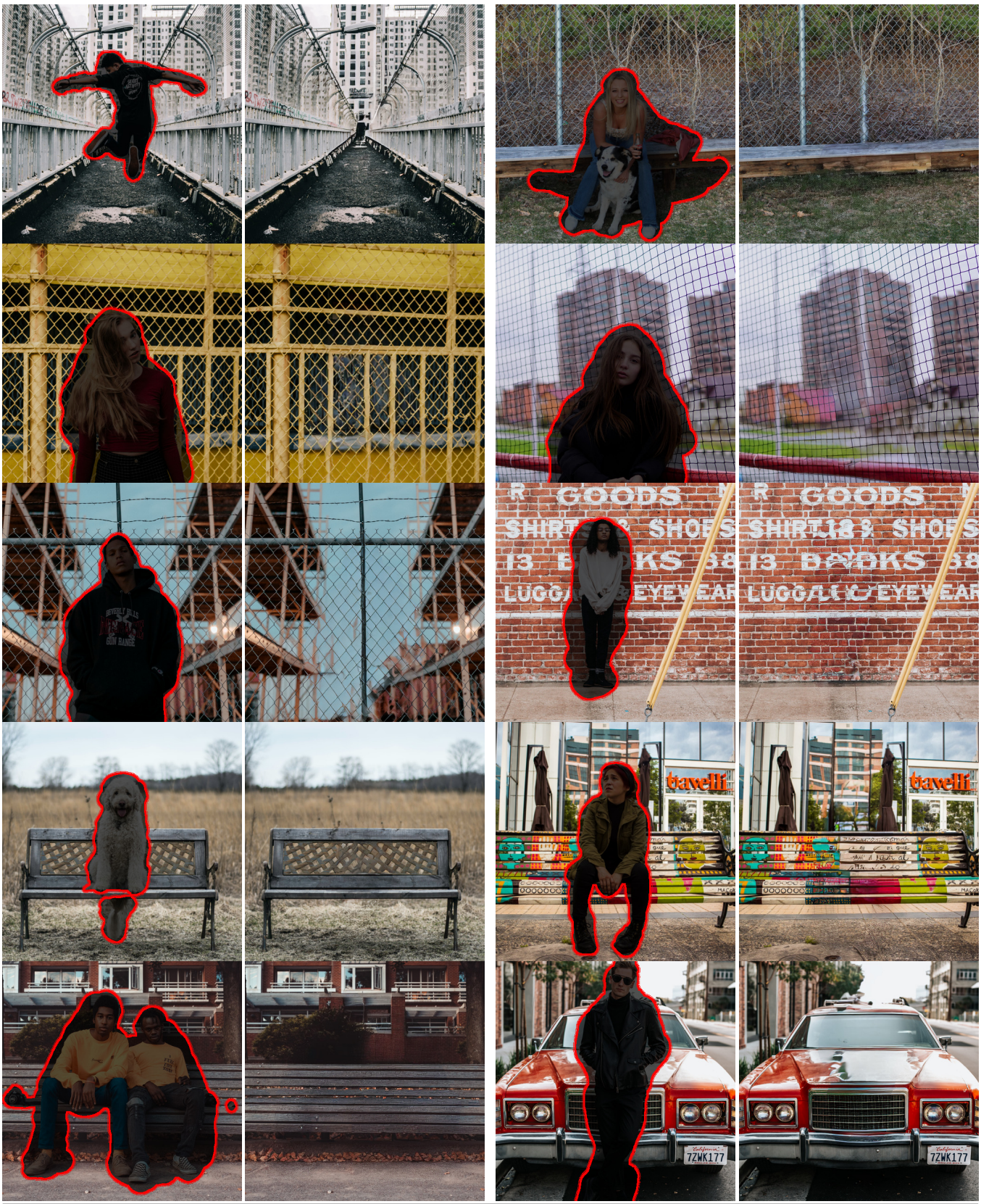

图22. 如图11所示的物体移除更多定性结果。

Semantic Synthesis on Flickr-Landscapes [23] (512^2 finetuning)

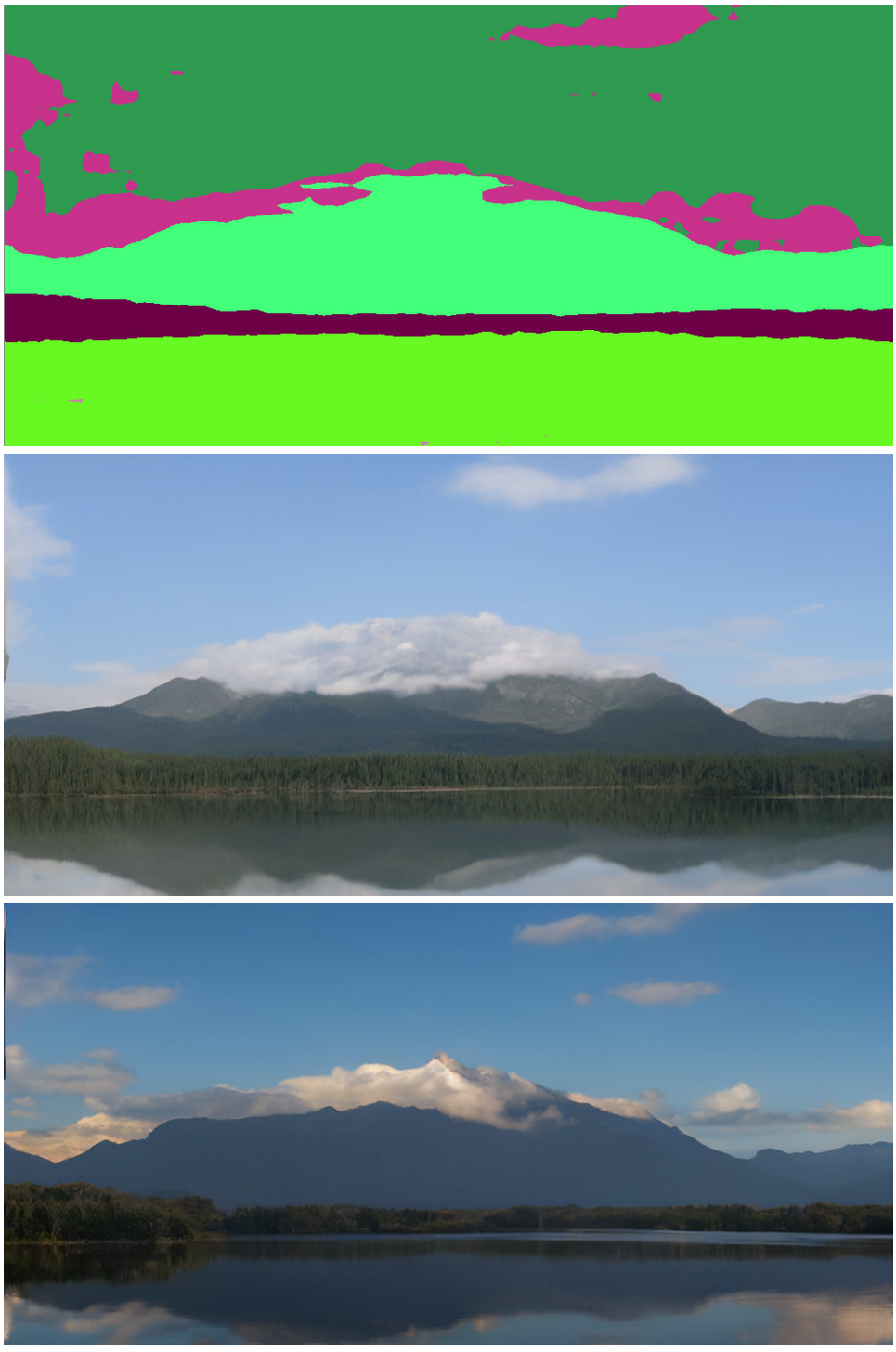

Figure 23. Convolutional samples from the semantic landscapes model as in Sec. 4.3.2, finetuned on 512^2 images.

Semantic Synthesis on Flickr-Landscapes [23] (512^2 finetuning)

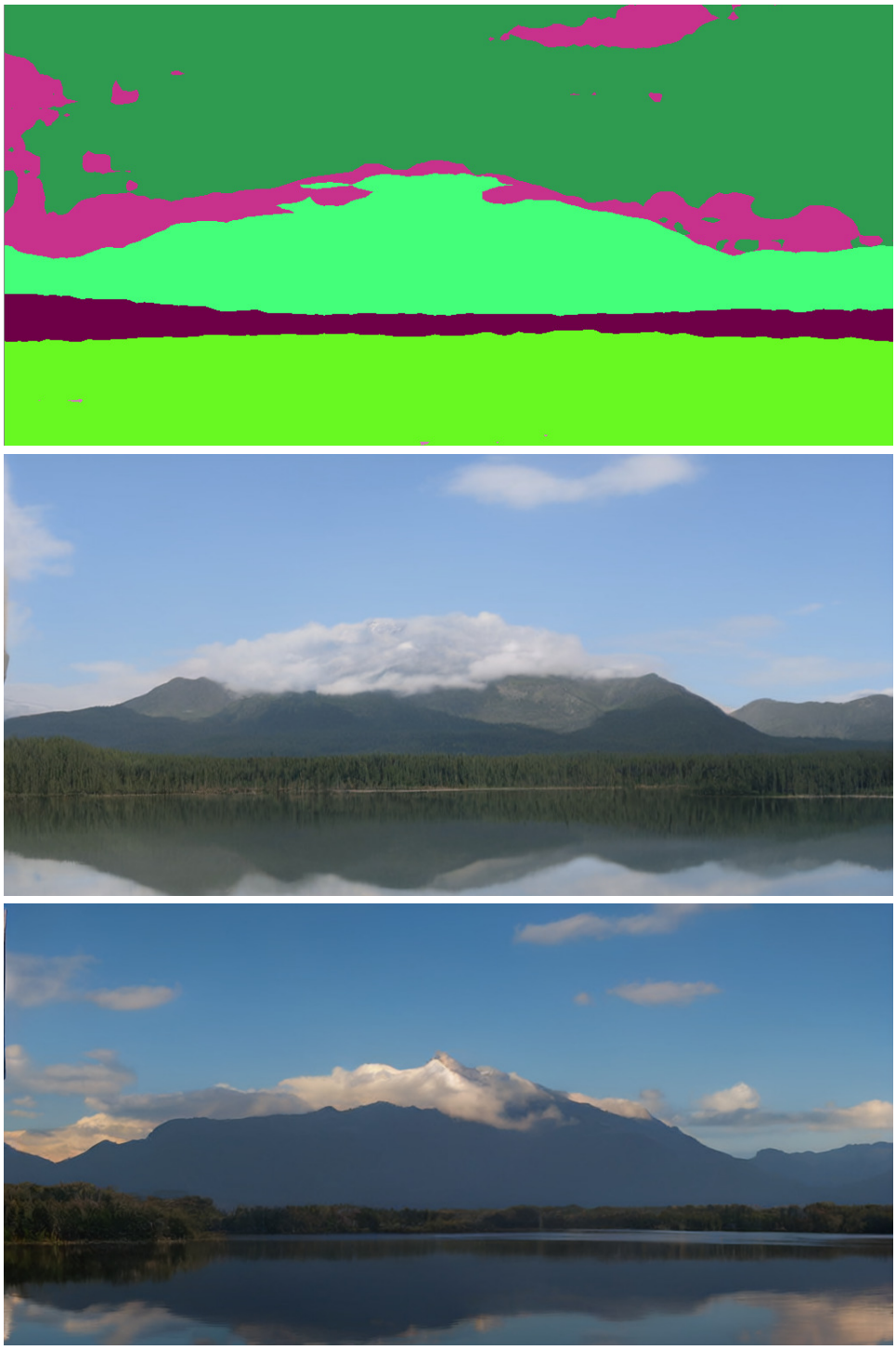

图23. 如第4.3.2节所述，在512^2图像上微调的语义景观模型的卷积样本。

Figure 24. A *LDM* trained on 256^2 resolution can generalize to larger resolution for spatially conditioned tasks such as semantic synthesis of landscape images. See Sec. 4.3.2.

图24. 在256^2分辨率上训练的*LDM*能够泛化至更高分辨率，适用于空间条件任务（如景观图像的语义合成）。详见第4.3.2节。

Semantic Synthesis on Flickr-Landscapes [23]

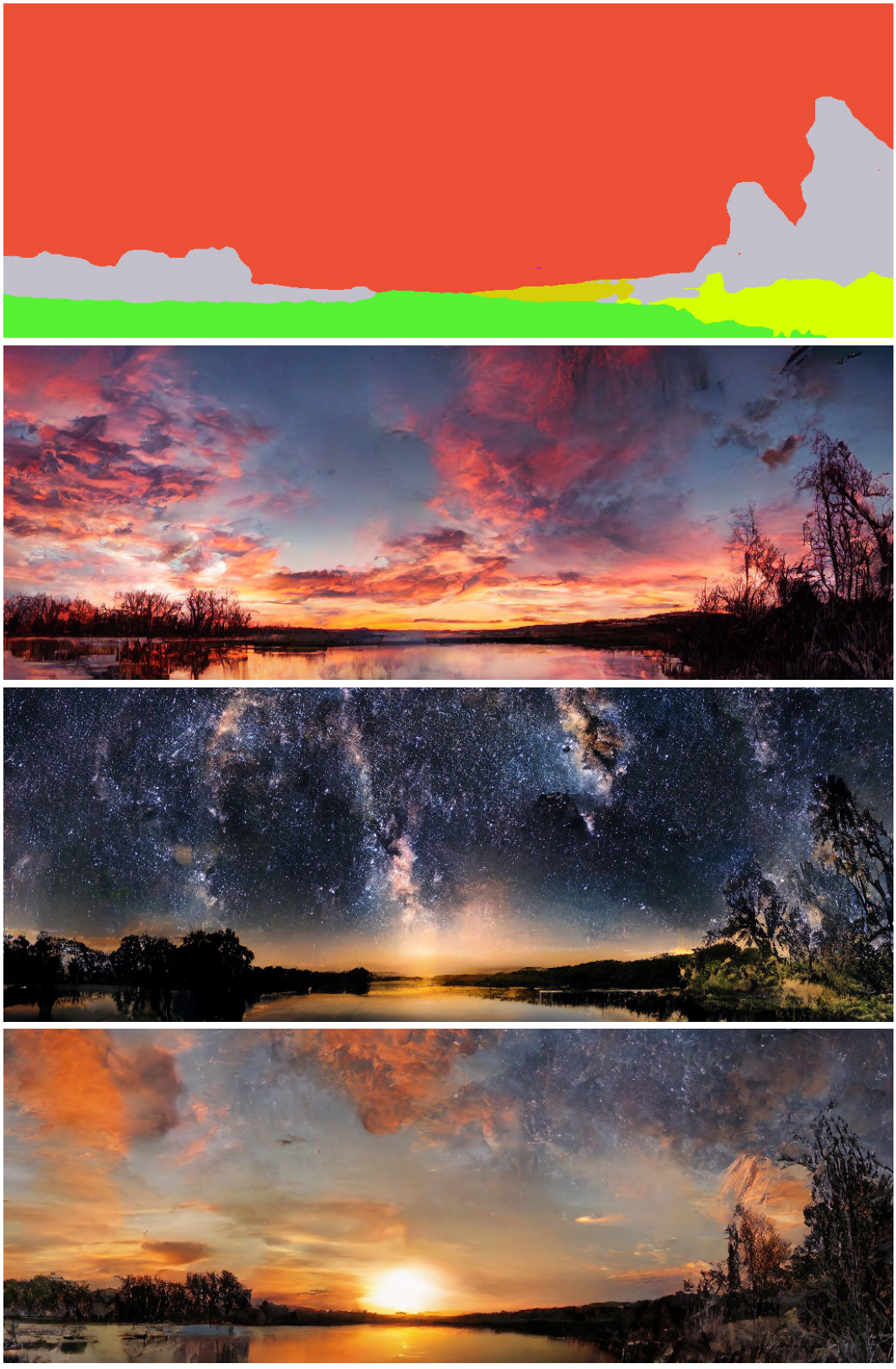

Figure 25. When provided a semantic map as conditioning, our *LDMs* generalize to substantially larger resolutions than those seen during training. Although this model was trained on inputs of size 256^2 it can be used to create high-resolution samples as the ones shown here, which are of resolution 1024×384.

Flickr-Landscapes上的语义合成[23]

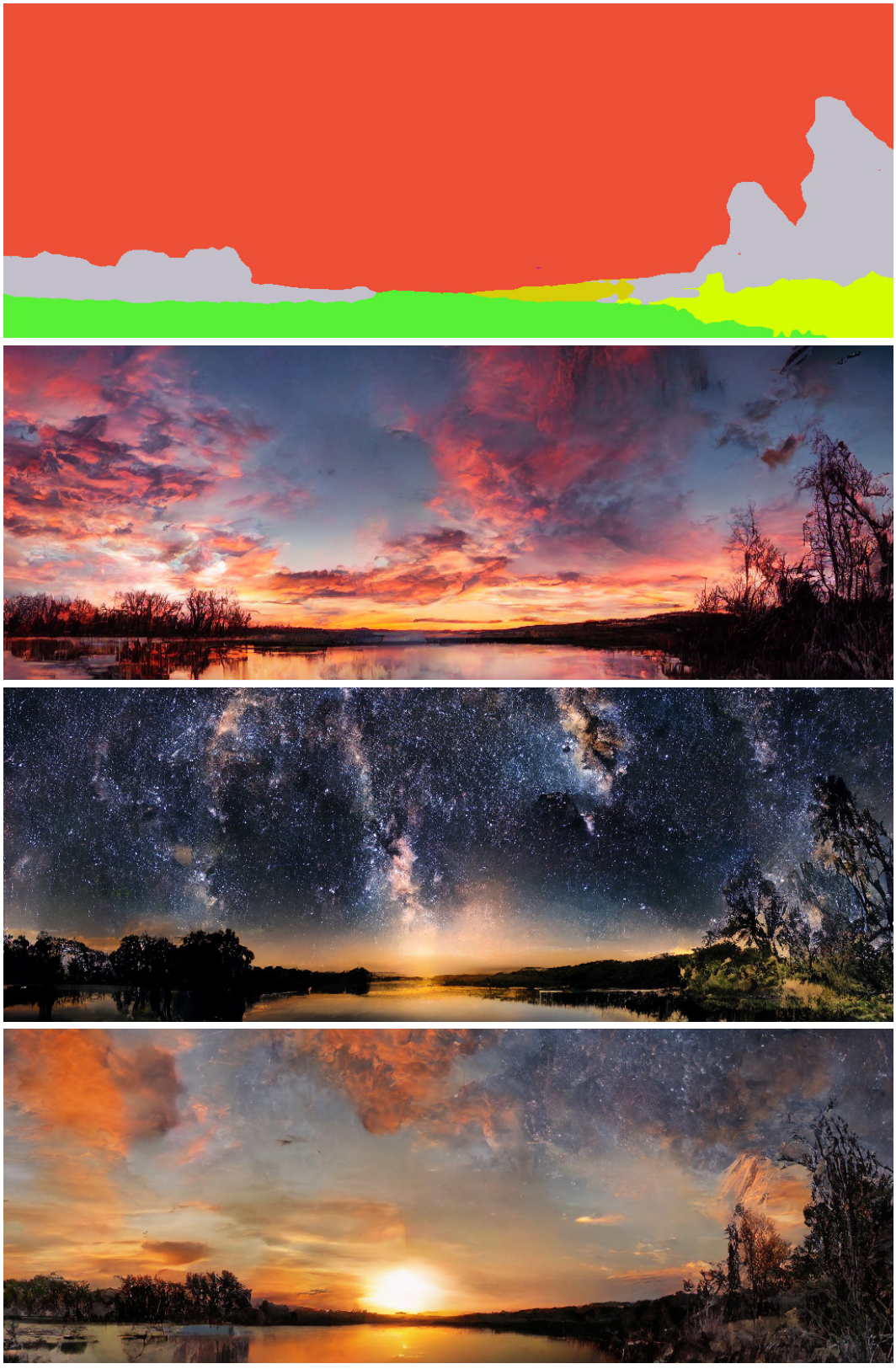

图25. 当提供语义地图作为条件时，我们的*LDMs*能够泛化到远大于训练所见的分辨率。尽管该模型是在256^2尺寸的输入上训练的，但它可用于生成如图所示的高分辨率样本，其分辨率达到1024×384。36

Random class conditional samples on the ImageNet dataset

Figure 26. Random samples from *LDM-4* trained on the ImageNet dataset. Sampled with classifier-free guidance [32] scale $s = 5.0$ and 200 DDIM steps with $\eta = 1.0$.

Random class conditional samples on the ImageNet dataset

图26. 在ImageNet数据集上训练的*LDM-4*随机样本。使用无分类器引导[32]尺度s =5.0和200步DDIM采样，η =1.0。

Random class conditional samples on the ImageNet dataset

Figure 27. Random samples from *LDM-4* trained on the ImageNet dataset. Sampled with classifier-free guidance [32] scale $s = 3.0$ and 200 DDIM steps with $\eta = 1.0$.

Random class conditional samples on the ImageNet dataset

图27. 在ImageNet数据集上训练的*LDM-4*随机样本。使用无分类器引导[32]采样，尺度为s =3.0，采用200步DDIM采样，η =1.0
。

Random samples on the CelebA-HQ dataset

Figure 28. Random samples of our best performing model *LDM-4* on the CelebA-HQ dataset. Sampled with 500 DDIM steps and $\eta = 0$ (FID = 5.15).

Random samples on the CelebA-HQ dataset

图28. 我们在CelebA-HQ数据集上表现最佳的模型*LDM-4*的随机样本。使用500步DDIM采样，η =为0（FID =为5.15）。

Random samples on the FFHQ dataset

Figure 29. Random samples of our best performing model *LDM-4* on the FFHQ dataset. Sampled with 200 DDIM steps and $\eta = 1$ (FID = 4.98).

Random samples on the FFHQ dataset

图29. 我们在FFHQ数据集上表现最佳的模型*LDM-4*的随机样本。使用200步DDIM采样，η =为1（FID=为4.98）。

Random samples on the LSUN-Churches dataset

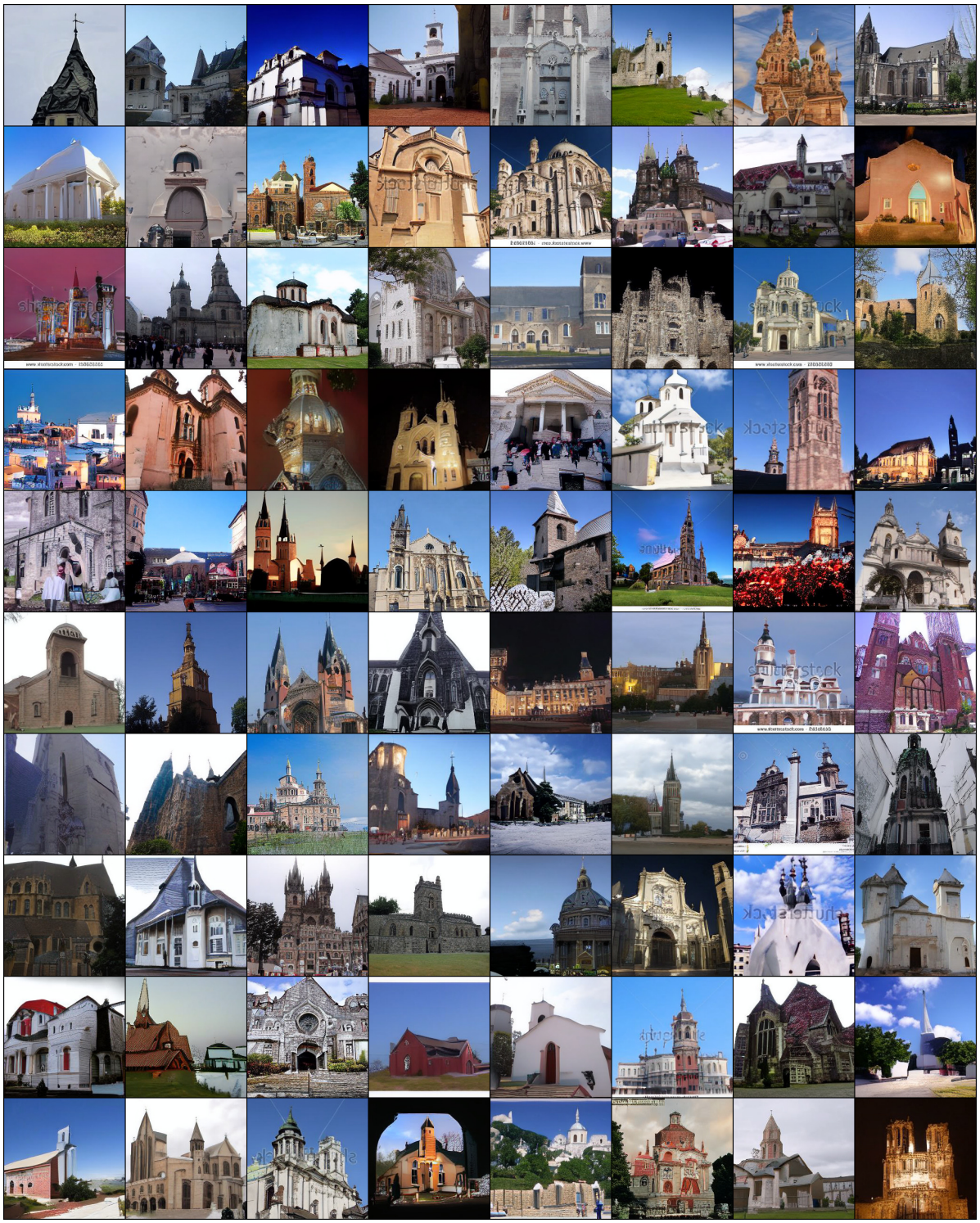

Figure 30. Random samples of our best performing model *LDM-8* on the LSUN-Churches dataset. Sampled with 200 DDIM steps and $\eta = 0$ (FID = 4.48).

Random samples on the LSUN-Churches dataset

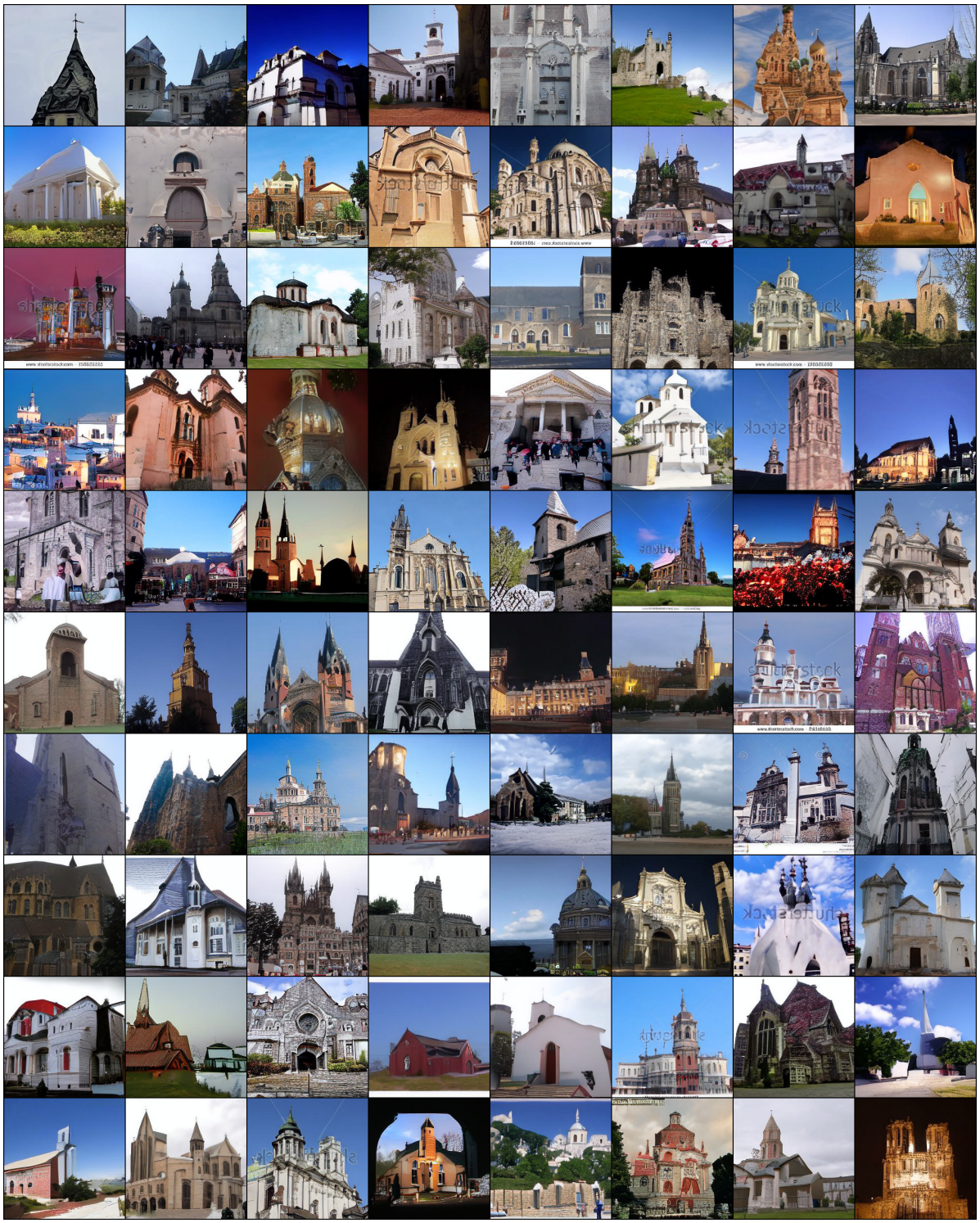

图30. 我们在LSUN-Churches数据集上表现最佳的模型*LDM-8*的随机样本。使用200步DDIM采样，η =为0（FID=为4.48）。

Random samples on the LSUN-Bedrooms dataset

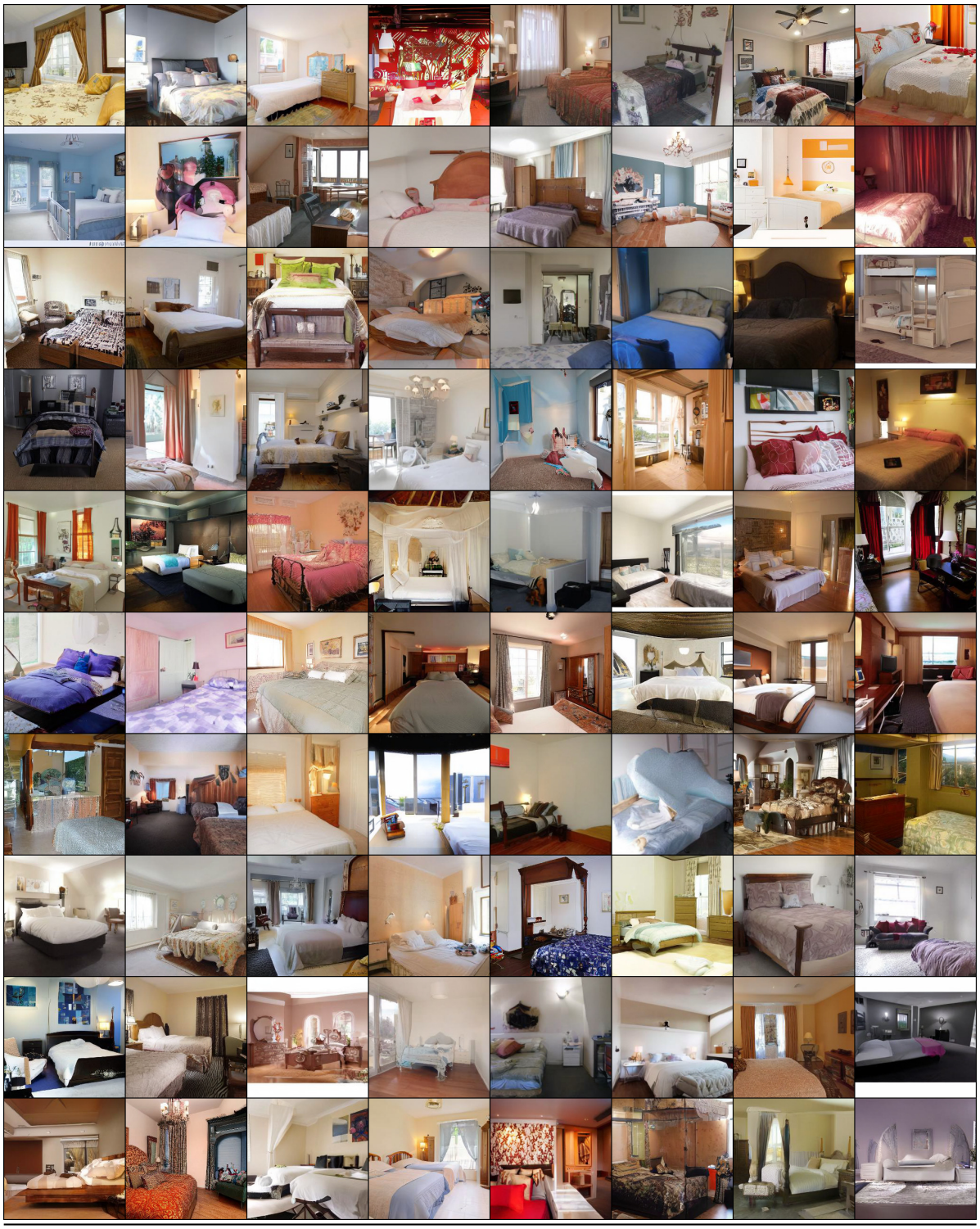

Figure 31. Random samples of our best performing model *LDM-4* on the LSUN-Bedrooms dataset. Sampled with 200 DDIM steps and $\eta = 1$ (FID = 2.95).

Random samples on the LSUN-Bedrooms dataset

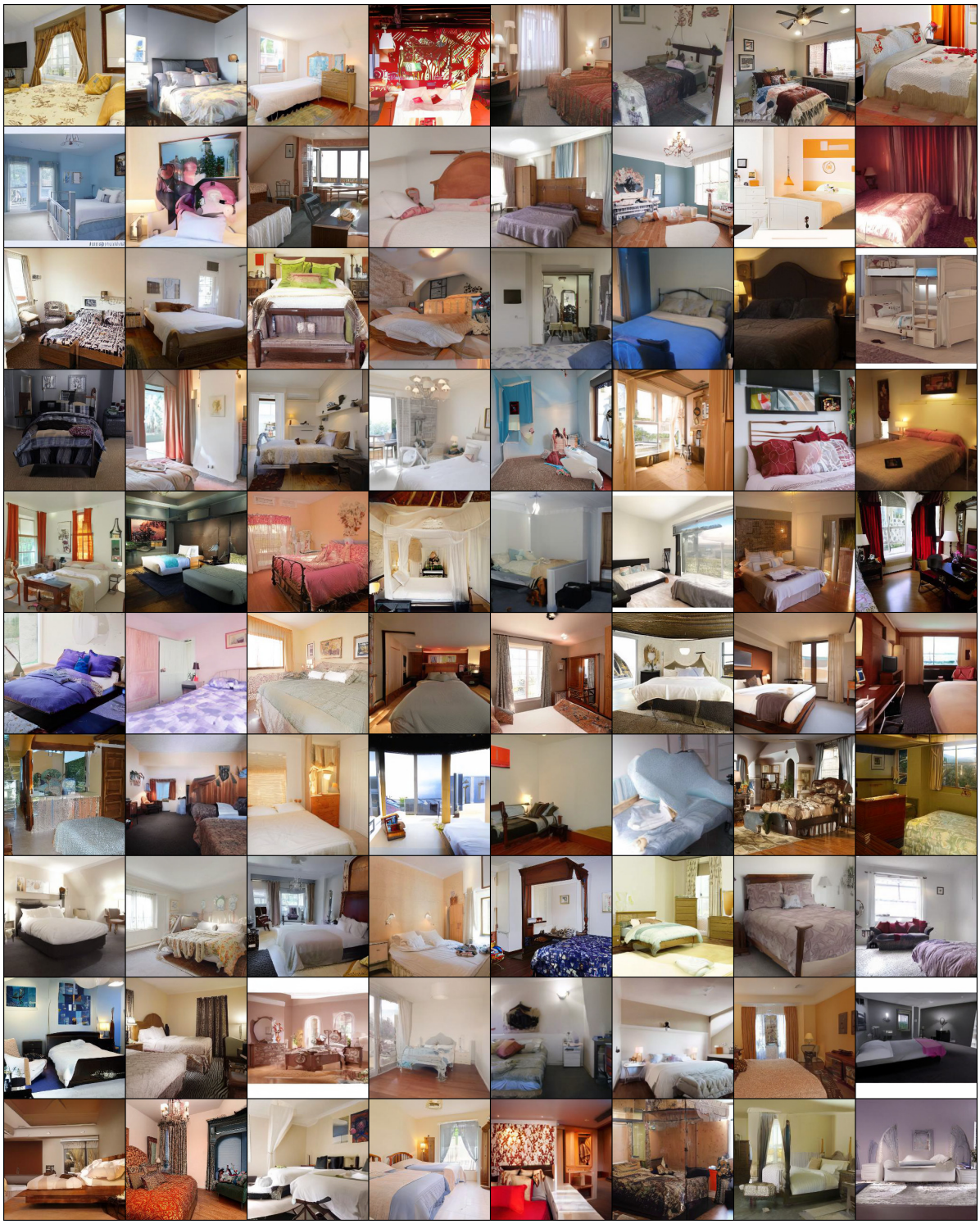

图31. 我们在LSUN卧室数据集上表现最佳的模型*LDM-4*的随机样本。使用200步DDIM采样，η =为1（FID =为2.95）。

Nearest Neighbors on the CelebA-HQ dataset

Figure 32. Nearest neighbors of our best CelebA-HQ model, computed in the feature space of a VGG-16 [79]. The leftmost sample is from our model. The remaining samples in each row are its 10 nearest neighbors.

Nearest Neighbors on the CelebA-HQ dataset

图32. 我们最佳CelebA-HQ模型的最近邻，在VGG-16 [79]的特征空间中计算。最左侧样本来自我们的模型。每行其余样本是其10个最近邻。

Nearest Neighbors on the FFHQ dataset

Figure 33. Nearest neighbors of our best FFHQ model, computed in the feature space of a VGG-16 [79]. The leftmost sample is from our model. The remaining samples in each row are its 10 nearest neighbors.

Nearest Neighbors on the FFHQ dataset

图33. 我们最佳FFHQ模型的最近邻样本，在VGG-16 [79]的特征空间中计算得出。最左侧样本来自我们的模型，每行其余样本为其10个最近邻。

Nearest Neighbors on the LSUN-Churches dataset

Figure 34. Nearest neighbors of our best LSUN-Churches model, computed in the feature space of a VGG-16 [79]. The leftmost sample is from our model. The remaining samples in each row are its 10 nearest neighbors.

Nearest Neighbors on the LSUN-Churches dataset

图34. 我们最佳LSUN教堂模型在VGG-16特征空间中的最近邻样本[79]。最左侧样本来自我们的模型，每行其余样本为其10个最近邻。