InstructPix2Pix：学习遵循图像编辑指令

InstructPix2Pix: Learning to Follow Image Editing Instructions

Tim Brooks*　　Aleksander Holynski*　　Alexei A. Efros

University of California, Berkeley

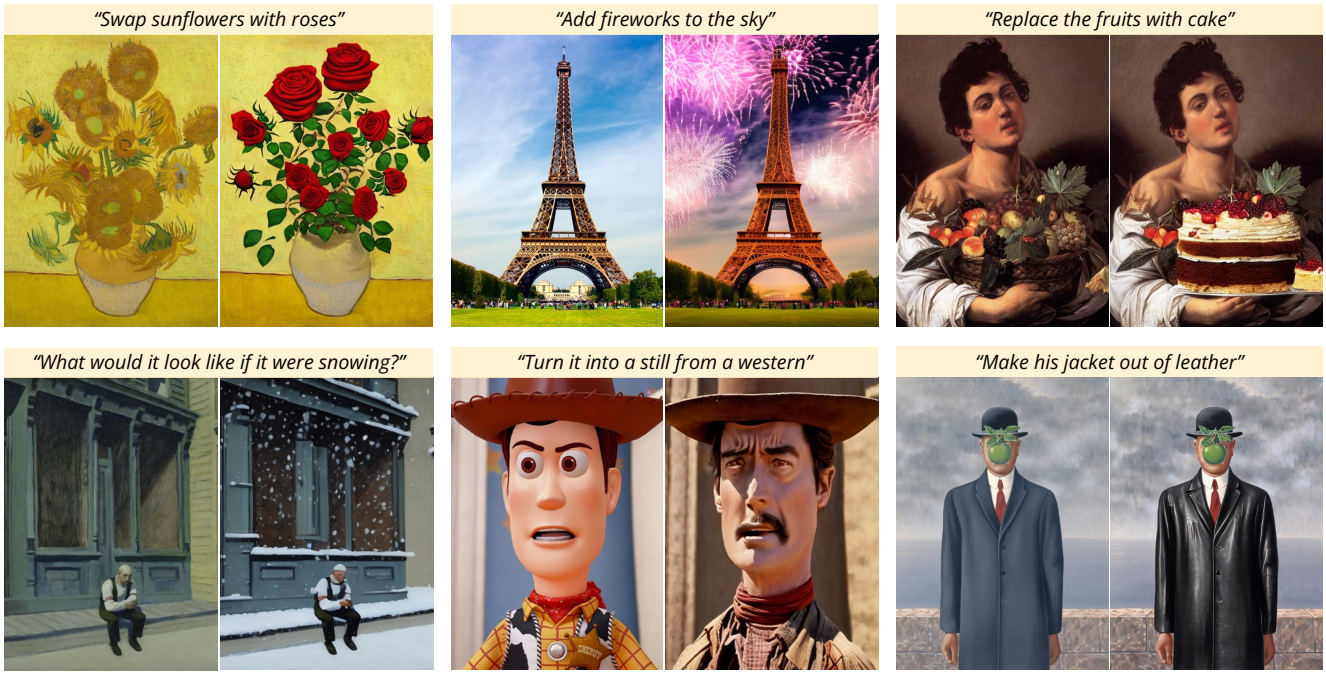


图 1. Given **an image** and **an instruction** for how to edit that image, our model performs the appropriate edit. Our model does not require full descriptions for the input or output image, and edits images in the forward pass without per-example inversion or fine-tuning. 给定**一张图像**和**编辑指令**，我们的模型执行相应的编辑操作。模型无需为输入或输出图像提供完整描述，可在前向传播中编辑图像，无需逐样本反演或微调。

摘要

We propose a method for editing images from human instructions: given an input image and a written instruction that tells the model what to do, our model follows these instructions to edit the image. To obtain training data for this problem, we combine the knowledge of two large pretrained models—a language model (GPT-3) and a text-to-image model (Stable Diffusion)—to generate a large dataset of image editing examples. Our conditional diffusion model, InstructPix2Pix, is trained on our generated data, and generalizes to real images and user-written instructions at inference time. Since it performs edits in the forward pass and does not require per-example fine-tuning or inversion, our model edits images quickly, in a matter of seconds. We show compelling editing results for a diverse collection of input images and written instructions.

我们提出一种基于人类指令的图像编辑方法。给定输入图像和编辑指令，模型遵循指令对图像进行编辑。为获取训练数据，我们结合两个大型预训练模型的知识——语言模型（*GPT-3*）和文本到图像模型（*Stable Diffusion*）——来生成大规模图像编辑示例数据集。我们的条件扩散模型 *InstructPix2Pix* 在生成的数据上训

练，可泛化到真实图像和用户指令。模型在前向传播中执行编辑，无需逐样本微调或反演，可在数秒内完成图像编辑。我们展示了对多样化输入图像和用户指令的显著编辑效果。

1. 引言（Introduction）

We present a method for teaching a generative model to follow human-written instructions for image editing. Since training data for this task is difficult to acquire at scale, we propose an approach for generating a paired dataset that combines multiple large models pretrained on different modalities: a large language model (GPT-3 [7]) and a text-to-image model (Stable Diffusion [52]). These two models capture complementary knowledge about language and images that can be combined to create paired training data for a task spanning both modalities.

我们提出一种方法，教导生成模型遵循人类指令进行图像编辑。由于此类任务的配对训练数据难以大规模获取，我们提出一种方法：结合在不同模态上预训练的多个大型模型——大型语言模型（GPT-3 [7]）和文本到图像模型（Stable Diffusion [52]）——来生成配对数据集。这两个模型各自捕捉语言和图像的互补知识，相结合可为跨越两个模态的任务创建配对训练数据。

Using our generated paired data, we train a conditional diffusion model that, given an input image and a text instruction for how to edit it, generates the edited image. Our model directly performs the image edit in the forward pass, and does not require any additional example images, full descriptions of the input/output images, or per-example finetuning. Despite being trained entirely on synthetic examples (i.e., both generated written instructions and generated imagery), our model achieves zero-shot generalization to both arbitrary *real* images and natural human-written instructions. Our model enables intuitive image editing that can follow human instructions to perform a diverse collection of edits: replacing objects, changing the style of an image, changing the setting, the artistic medium, among others. Selected examples can be found in Figure 1.

利用生成的配对数据，我们训练条件扩散模型。给定输入图像和编辑指令，模型生成编辑后的图像。模型在前向传播中直接执行编辑，无需额外示例图像、完整的输入/输出描述或逐样本微调。尽管仅在合成数据（生成的指令和生成的图像）上训练，模型可泛化到任意真实图像和自然用户指令。模型支持直观的图像编辑：替换物体、改变风格、改变场景、改变艺术风格等。示例见图 1。

2. 相关工作（Prior work）

组合大型预训练模型（Composing large pretrained models） Recent work has shown that large pretrained models can be combined to solve multimodal tasks that no one model can perform alone, such as image captioning and visual question answering (tasks that require the knowledge of both a large language model and a text-image model). Techniques for combining pretrained models include joint finetuning on a new task [4, 34, 41, 68], communication through prompting [63, 70], composing probability distributions of energy-based models [11, 38], guiding one model with feedback from another [62], and iterative optimization [35]. Our method is similar to prior work in that it leverages the complementary abilities of two pretrained models—GPT-3 [7]) and Stable Diffusion [52]—but differs in that we use these models to generate paired multi-modal training data.

最近的研究表明，大型预训练模型可以结合起来解决单个模型无法完成的多模态任务，如图像描述和视觉问答（需要大型语言模型和文本-图像模型的知识）。结合预训练模型的技术包括在新任务上联合微调 [4, 34, 41, 68]、通过提示进行通信 [63, 70]、组合基于能量的模型的概率分布 [11, 38]、用另一模型的反馈引导某个模型 [62] 以及迭代优化 [35]。我们的方法与之前工作相似，都利用两个预训练模型的互补能力——GPT-3 [7] 和 Stable Diffusion [52]——但不同之处在于，我们用这些模型生成配对的多模态训练数据。

基于扩散的生成模型（Diffusion-based generative models） Recent advances in diffusion models [60] have enabled state-of-the-art image synthesis [10, 18, 19, 54, 56, 61] as well as generative models of other modalities such as video [21, 59], audio [31], text [36] and network parameters [46]. Recent text-to-image diffusion models [42, 49, 52, 55] have shown to generate realistic images from arbitrary text captions.

扩散模型的最近进展 [60] 实现了最先进的图像合成 [10, 18, 19, 54, 56, 61]，以及视频 [21, 59]、音频 [31]、文本 [36] 和网络参数 [46] 等其他模态的生成模型。最近的文本到图像扩散模型 [42, 49, 52, 55] 可从任意文本

* 表示同等贡献

更多结果见项目主页：timothybrooks.com/instruct-pix2pix

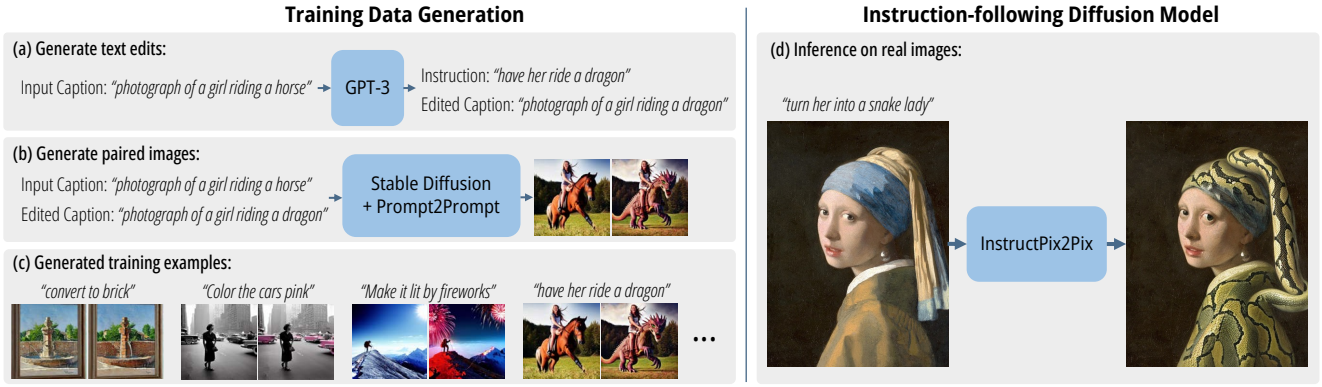


图 2. Our method consists of two parts: generating an image editing dataset, and training a diffusion model on that dataset. (a) We first use a finetuned GPT-3 to generate instructions and edited captions. (b) We then use StableDiffusion [52] in combination with Prompt-to-Prompt [17] to generate pairs of images from pairs of captions. We use this procedure to create a dataset (c) of over 450,000 training examples. (d) Finally, our InstructPix2Pix diffusion model is trained on our generated data to edit images from instructions. At inference time, our model generalizes to edit real images from human-written instructions. 我们的方法分为两部分：生成图像编辑数据集，以及在该数据集上训练扩散模型。(a) 首先利用微调的 GPT-3 生成指令和编辑后的描述文本。(b) 然后使用 Stable Diffusion [52] 结合 Prompt-to-Prompt [17] 从成对的描述文本生成成对的图像。利用这一过程生成包含超过 45 万个训练样例的数据集 (c)。(d) 最后，我们的 InstructPix2Pix 扩散模型在生成的数据上训练，用于从指令编辑图像。推理时，模型可泛化到编辑真实图像和用户指令。

描述生成真实感图像。

图像编辑的生成模型（Generative models for image editing） Image editing models traditionally targeted a single editing task such as style transfer [15, 16] or translation between image domains [22, 24, 37, 43, 72]. Numerous editing approaches invert [1–3, 12] or encode [8, 51, 64] images into a latent space (e.g., StyleGAN [26, 27]) where they can be edited by manipulating latent vectors. Recent models have leveraged CLIP [48] embeddings to guide image editing using text [5, 9, 14, 29, 32, 42, 45, 71]. We compare with one of these methods, Text2Live [6], an editing method that optimizes for an additive image layer that maximizes a CLIP similarity objective.

图像编辑模型传统上针对单一编辑任务，如风格迁移 [15, 16] 或图像域之间的转换 [22, 24, 37, 43, 72]。许多编辑方法将图像反演 [1–3, 12] 或编码 [8, 51, 64] 到潜在空间（如 StyleGAN [26, 27]），在此空间中可通过操纵潜在向量进行编辑。最近的模型利用 CLIP [48] 嵌入来指导文本-引导的图像编辑 [5, 9, 14, 29, 32, 42, 45, 71]。我们与其中一种方法（Text2Live [6]）进行了比较，该方法通过优化加法图像层来最大化 CLIP 相似性目标。

Recent works have used pretrained text-to-image diffusion models for image editing [5, 17, 28, 39, 49]. While some text-to-image models natively have the ability to edit images (e.g., DALLE-2 can create variations of images, inpaint regions, and manipulate the CLIP embedding [49]), using these models for *targeted* editing is non-trivial, because in most cases they offer no guarantees that similar text prompts will yield similar images. Recent work by Hertz *et al.* [17] tackles this issue with Prompt-to-Prompt, a method for assimilating the generated images for similar text prompts, such that isolated edits can be made to a generated image. We use this method in generating training data. To edit non-generated (i.e., real) imagery, SDEdit [39] uses a pretrained model to noise and denoise an input image with a new target prompt. We compare with SDEdit as a baseline. Other recent works perform local inpainting given a caption and user-drawn mask [5, 49], generate new images of a specific object or concept learned from a small collection of images [13, 53], or perform editing by inverting (and fine-tuning) a single image, and subsequently regenerating with a new text description [28]. In contrast to these approaches, our model takes only a single image and an instruction for how to edit that image (i.e., not a full description of any image), and performs the edit directly in the forward pass without need for a user-drawn mask, additional images, or per-example inversion or fine-tuning.

最近的工作使用预训练的文本到图像扩散模型进行图像编辑 [5, 17, 28, 39, 49]。虽然一些文本到图像模型本身具有编辑图像的能力（例如 DALLE-2 可生成图像变化、修复区域及操纵 CLIP 嵌入 [49]），但用于针对性编辑并非平凡，因为它们通常无法保证相似文本提示会产生相似图像。Hertz 等人 [17] 的最近工作通过 Prompt-to-Prompt 方法解决此问题，该方法使相似文本提示生成的图像相一致，允许对生成的图像进行孤立编辑。我们在生成训练数据时采用了该方法。对于真实图像编辑，SDEdit [39] 使用预训练模型对输入图像加噪和去噪，以符合新目标提示。我们以 SDEdit 作为基线进行比较。其他最近的工作在给定描述文本和用户绘制掩膜的条件下进行局部补全 [5, 49]、从小图像集学习特定物体或概念随后生成新图像 [13, 53]，或通过反演和微调单个图像、随后用新文本描述重新生成来编辑 [28]。相比之下，我们的模型仅需输入图像和编辑指令（而非任何图像的完整描述），可在前向传播中直接执行编辑，无需用户掩膜、额外图像或逐样本反演/微调。

学习遵循指令（Learning to follow instructions） Our method differs from existing text-based image editing works [6, 13, 17, 28, 39, 53] in that it enables editing from *instructions* that tell the model what action to perform, as opposed to text labels, captions or descriptions of input/output images. A key benefit of following editing instructions is that the user can just tell the model exactly what to do in natural written text. There is no need for the user to provide extra information, such as example images or descriptions of visual content that remains constant between the input and output images. Instructions are expressive, precise, and intuitive to write, allowing the user to easily isolate specific objects or visual attributes to change. Our goal to follow written image editing instructions is inspired by recent work teaching large language models to better follow human instructions for language tasks [40, 44, 69].

我们的方法与现有基于文本的图像编辑工作 [6, 13, 17, 28, 39, 53] 的不同之处在于，它允许从指令进行编辑，这些指令告诉模型要执行什么操作，而不是文本标签、描述文本或输入/输出图像的描述。遵循编辑指令的关键优势是用户可以用自然书面文本准确告诉模型要做什么。用户无需提供额外信息，如示例图像或描述输入和输出图像之间保持不变的视觉内容的描述。指令具有表现力、精确性且易于书写，允许用户轻松隔离特定物体或视觉属性进行修改。让模型遵循书面编辑指令，这一目标受到近期研究的启发：这些研究教大型语言模型更好地遵循语言任务中的人类指令 [40, 44, 69]。

使用生成模型生成训练数据（Training data generation with generative models） Deep models typically require large amounts of training data. Internet data collections are often suitable, but may not exist in the form necessary for supervision, e.g., paired data of particular modalities. As generative models continue to improve, there is growing interest in their use as a source of cheap and plentiful training data for downstream tasks [33, 47, 50, 58, 65, 66]. In this paper, we use two different off-the-shelf generative models (language, text-to-image) to produce training data for our editing model.

深度模型通常需要大量训练数据。互联网数据集通常适用，但可能不以监督所需的形式存在，例如特定模态的配对数据。随着生成模型的持续改进，人们日益关注将其作为下游任务廉价而丰富的训练数据来源 [33, 47, 50, 58, 65, 66]。本文使用两个不同的现成生成模型（语言、文本到图像）来为编辑模型生成训练数据。

3. 方法（Method）

We treat instruction-based image editing as a supervised learning problem: (1) first, we generate a paired training dataset of text editing instructions and images before/after the edit (Sec. 3.1, Fig. 2a-c), then (2) we train an image editing diffusion model on this generated dataset (Sec. 3.2, Fig 2d). Despite being trained with generated images and editing instructions, our model is able to generalize to editing *real* images using arbitrary human-written instructions. See Fig. 2 for an overview of our method.

我们将基于指令的图像编辑视为监督学习问题：（1）首先生成成对的训练数据集，包含文本编辑指令与编辑前后的图像（第 3.1 节，图 2a-c）；（2）在此数据集上训练图像编辑扩散模型（第 3.2 节，图 2d）。尽管采用生成的图像和指令进行训练，我们的模型能够泛化到使用任意人工编写的指令编辑真实图像。图 2 展示了方法的整体概览。

3.1. 生成多模态训练数据集（Generating a Multimodal Training Dataset）

We combine the abilities of two large-scale pretrained models that operate on different modalities—a large language model [7] and a text-to-image model [52]—to gener-

ate a multi-modal training dataset containing text editing instructions and the corresponding images before and after the edit. In the following two sections, we describe in detail the two steps of this process. In Section 3.1.1, we describe the process of fine-tuning GPT-3 [7] to generate a collection of text edits: given a prompt describing an image, produce a text instruction describing a change to be made and a prompt describing the image after that change (Figure 2a). Then, in Section 3.1.2, we describe the process of converting the two text prompts (i.e., before and after the edit) into a pair of corresponding images using a text-to-image model [52] (Figure 2b).

我们结合两个大规模预训练模型的能力——分别在不同模态上运作的大型语言模型 [7] 和文本到图像模型 [52]——来生成包含文本编辑指令及其对应编辑前后图像的多模态训练数据集。在随后两小节中，我们详细描述此过程的两个步骤。第 3.1.1 节描述微调 GPT-3 [7] 的过程，生成一组文本编辑：给定描述图像的文本提示，生成描述所需修改的指令和描述修改后图像的文本提示（图 2a）。第 3.1.2 节则描述将这两个文本提示（编辑前后）转换为配对图像的过程，使用文本到图像模型 [52]（图 2b）。

3.1.1 生成指令与配对描述文本（Generating Instructions and Paired Captions）

We first operate entirely in the text domain, where we leverage a large language model to take in image captions and produce editing instructions and the resulting text captions after the edit. For example, as shown in Figure 2a, provided the input caption *"photograph of a girl riding a horse"*, our language model can generate both a plausible edit instruction *"have her ride a dragon"* and an appropriately modified output caption *"photograph of a girl riding a dragon"*. Operating in the text domain enables us to generate a large and diverse collection of edits, while maintaining correspondence between the image changes and text instructions.

我们首先完全在文本域中操作，利用大型语言模型将图像描述文本转化为编辑指令和修改后的描述文本。例如，图 2a 所示，给定输入描述文本 "照片：女孩骑马"，我们的语言模型可生成合理的编辑指令 "让她骑龙" 和相应修改的输出描述文本 "照片：女孩骑龙"。在文本域操作使我们能生成大规模、多样化的编辑集合，同时保持图像变化与文本指令间的对应性。

Our model is trained by finetuning GPT-3 on a relatively small human-written dataset of editing triplets: (1) input captions, (2) edit instructions, (3) output captions. To produce the fine-tuning dataset, we sampled 700 input captions from the LAION-Aesthetics V2 6.5+ [57] dataset and manually wrote instructions and output captions. See Table 1a for examples of our written instructions and output captions. Using this data, we fine-tuned the GPT-3 Davinci model for a single epoch using the default training parameters.

我们通过在相对较小的人工标注编辑三元组数据集上微调 GPT-3 来训练模型，数据包括：（1）输入描述文本、（2）编辑指令、（3）输出描述文本。为生成微调数据集，我们从 LAION-Aesthetics V2 6.5+ [57] 数据集中采样 700 个输入描述文本，并手工编写指令和输出描述文本。表 1a 展示了我们编写的指令和输出描述文本的示例。基于此数据，我们使用默认训练参数对 GPT-3 Davinci 模型进行了一轮微调。

Benefiting from GPT-3's immense knowledge and ability to generalize, our finetuned model is able to generate creative yet sensible instructions and captions. See Table 1b for example GPT-3 generated data. Our dataset is created by generating a large number of edits and output captions using this trained model, where the input captions are real image captions from LAION-Aesthetics (excluding samples with duplicate captions or duplicate image URLs). We chose the LAION dataset due to its large size, diversity of content (including references to proper nouns and popular culture), and variety of mediums (photographs, paintings, digital artwork). A potential drawback of LAION is that it is quite noisy and contains a number of nonsensical or undescriptive captions—however, we found that dataset noise is mitigated through a combination of dataset filtering (Section 3.1.2) and classifier-free guidance (Section 3.2.1). Our final corpus of generated instructions and captions consists of 454,445 examples.

由于 GPT-3 具备丰富的知识和泛化能力，我们的微调模型能生成既具创意又合理的指令和描述文本。表 1b 展示了 GPT-3 生成数据的示例。数据集通过使用训练好的模型生成大量编辑和输出描述文本而构建，其中输入描述文本为 LAION-Aesthetics 中真实的图像描述（排除重复描述或重复图像 URL 的样本）。我们选择 LAION 数据集源于其规模庞大、内容多样（包含专有名词和流行文化的引用）和介质丰富（照片、绘画、数字艺术作品）。LAION 的缺点是噪声较大，包

	输入 LAION 描述文本	编辑指令	编辑后描述文本
人工编写（700 条编辑）	*Yefim Volkov, Misty Morning*	*make it afternoon*	*Yefim Volkov, Misty Afternoon*
	girl with horse at sunset	*change the background to a city*	*girl with horse at sunset in front of city*
	painting-of-forest-and-pond	*Without the water.*	*painting-of-forest*
	...	...	...
GPT-3 生成（>450,000 条编辑）	*Alex Hill, Original oil painting on canvas, Moonlight Bay*	*in the style of a coloring book*	*Alex Hill, Original coloring book illustration, Moonlight Bay*
	The great elf city of Rivendell, sitting atop a waterfall as cascades of water spill around it	*Add a giant red dragon*	*The great elf city of Rivendell, sitting atop a waterfall as cascades of water spill around it with a giant red dragon flying overhead*
	Kate Hudson arriving at the Golden Globes 2015	*make her look like a zombie*	*Zombie Kate Hudson arriving at the Golden Globes 2015*
	...	...	...

表 1. We label a small text dataset, finetune GPT-3, and use that finetuned model to generate a large dataset of text triplets. As the input caption for both the labeled and generated examples, we use real image captions from LAION. Highlighted text is generated by GPT-3. 我们标注小规模文本数据集、微调 GPT-3，并使用微调后的模型生成大规模文本三元组数据集。对于标注和生成的样例，我们均使用 LAION 中真实的图像描述文本作为输入。高亮文本由 GPT-3 生成。

(a) Without Prompt-to-Prompt. 不使用 Prompt-to-Prompt。

(b) With Prompt-to-Prompt. 使用 Prompt-to-Prompt。

图 3. Pair of images generated using StableDiffusion [52] with and without Prompt-to-Prompt [17]. For both, the corresponding captions are *"photograph of a girl riding a horse"* and *"photograph of a girl riding a dragon"*.

含许多无意义或不具描述性的描述文本。然而，我们发现通过结合数据集筛选（第 3.1.2 节）和无分类器引导（第 3.2.1 节），可以有效缓解这一噪声问题。最终生成的指令和描述文本语料库包含 454,445 个示例。

3.1.2 从配对描述文本生成配对图像（Generating Paired Images from Paired Captions）

Next, we use a pretrained text-to-image model to transform a pair of captions (referring to the image before and after the edit) into a pair of images. One challenge in turning a pair of captions into a pair of corresponding images is that text-to-image models provide no guarantees about image consistency, even under very minor changes of the conditioning prompt. For example, two very similar prompts: *"a picture of a cat"* and *"a picture of a black cat"* may produce wildly different images of cats. This is unsuitable for our purposes, where we intend to use this paired data as supervision for training a model to edit images (and not produce a different random image). We therefore use Prompt-to-Prompt [17], a recent method aimed at encouraging multiple generations from a text-to-image diffusion model to be similar. This is done through borrowed cross attention weights in some number of denoising steps. Figure 3 shows a comparison of sampled images with and without Prompt-to-Prompt.

接下来，我们使用预训练的文本到图像模型将一对描述文本（分别代表编辑前后的图像）转换为一对图像。将一对描述文本转换为一对对应图像的难点在于，文本到图像模型不保证图像一致性，即使在极小的条件文本修改下也是如此。例如，两个非常相似的文本提示 "猫的图片" 和 "黑猫的图片" 可能产生截然不同的猫图。这对我们的目标不适用，我们希望将这些配对数据作为监督信号来训练图像编辑模型（而非生成不同的随机图像）。因此我们使用 Prompt-to-Prompt [17]，一种旨在使文本到图像扩散模型的多个生成结果相似

的最近方法。这通过借用部分去噪步骤中的交叉注意力权重来实现。图 3 展示了有无 Prompt-to-Prompt 方法的图像比较。

While this greatly helps assimilate generated images, different edits may require different amounts of change in image-space. For instance, changes of larger magnitude, such as those which change large-scale image structure (e.g., moving objects around, replacing with objects of different shapes), may require less similarity in the generated image pair. Fortunately, Prompt-to-Prompt has as a parameter that can control the similarity between the two images: the fraction of denoising steps p with shared attention weights. Unfortunately, identifying an optimal value of p from only the captions and edit text is difficult. We therefore generate 100 sample pairs of images per caption-pair, each with a random $p \sim \mathcal{U}(0.1, 0.9)$, and filter these samples by using a CLIP-based metric: the directional similarity in CLIP space as introduced by Gal *et al.* [14]. This metric measures the consistency of the change between the two images (in CLIP space) with the change between the two image captions. Performing this filtering not only helps maximize the diversity and quality of our image pairs, but also makes our data generation more robust to failures of Prompt-to-Prompt and Stable Diffusion.

虽然这有助于图像的对齐，不同的编辑可能需要在图像空间中不同程度的变化。例如，较大幅度的修改（如改变大尺度图像结构、移动物体或用不同形状的物体替换），在生成的图像对中可能需要较低的相似度。幸运的是，Prompt-to-Prompt 有一个参数可控制两个图像间的相似度：具有共享注意力权重的去噪步骤比例 p。不幸的是，仅从描述文本和编辑文本中识别最优 p 值很困难。因此，我们为每对描述文本生成 100 对图像样本，每个样本使用随机采样的 $p \sim \mathcal{U}(0.1, 0.9)$，并使用基于 CLIP 的指标进行筛选：使用 Gal 等人 [14] 引入的 CLIP 空间中的方向相似度。该指标度量两个图像间变化的一致性（在 CLIP 空间中）与两个图像描述文本间变化的一致性。此筛选不仅最大化了图像对的多样性和质量，还使我们的数据生成对 Prompt-to-Prompt 和 Stable Diffusion 故障更加鲁棒。

3.2. InstructPix2Pix

We use our generated training data to train a conditional diffusion model that edits images from written instructions. We base our model on Stable Diffusion, a large-scale text-to-image latent diffusion model.

我们使用生成的训练数据来训练一个条件扩散模型，该模型根据文本指令进行图像编辑。我们的模型基于 Stable Diffusion，一个大规模的文本到图像潜在扩散模型。

Diffusion models [60] learn to generate data samples through a sequence of denoising autoencoders that estimate the score [23] of a data distribution (a direction pointing toward higher density data). Latent diffusion [52] improves the efficiency and quality of diffusion models by operating in the latent space of a pretrained variational autoencoder [30] with encoder $\mathcal{E}$ and decoder $\mathcal{D}$. For an image x, the diffusion process adds noise to the encoded latent $z = \mathcal{E}(x)$ producing a noisy latent z_t where the noise level increases over timesteps $t \in T$. We learn a network ϵ_θ that predicts the noise added to the noisy latent z_t given image conditioning c_I and text instruction conditioning c_T. We minimize the following latent diffusion objective:

扩散模型 [60] 通过一系列去噪自编码器学习生成数据样本，这些自编码器估计数据分布的分数 [23]（指向更高密度数据的方向）。潜在扩散 [52] 通过在预训练变分自编码器 [30] 的潜在空间中操作来提高扩散模型的效率和质量，该自编码器具有编码器 $\mathcal{E}$ 和解码器 $\mathcal{D}$。对于图像 x，扩散过程向编码后的潜在变量 $z = \mathcal{E}(x)$ 中加入噪声，生成带噪潜在变量 z_t，其噪声水平随时间步 $t \in T$ 增加。我们学习一个网络 ϵ_θ，给定图像条件 c_I 和文本指令条件 c_T，预测加入到带噪潜在变量 z_t 中的噪声。我们最小化以下潜在扩散目标：

$$L = \mathbb{E}_{\mathcal{E}(x),\mathcal{E}(c_I),c_T,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, \mathcal{E}(c_I), c_T))\|_2^2\right] \tag{1}$$

Wang *et al.* [67] show that fine-tuning a large image diffusion models outperforms training a model from scratch for image translation tasks, especially when paired training data is limited. We therefore initialize the weights of our model with a pretrained Stable Diffusion checkpoint, leveraging its vast text-to-image generation capabilities. To support image conditioning, we add additional input channels to the first convolutional layer, concatenating z_t and $\mathcal{E}(c_I)$. All available weights of the diffusion model are initialized from the pretrained checkpoints, and weights that operate on the newly added input channels are initialized to zero. We reuse the same text conditioning mechanism that was originally intended for captions to instead take as input the text edit instruction c_T. Additional training

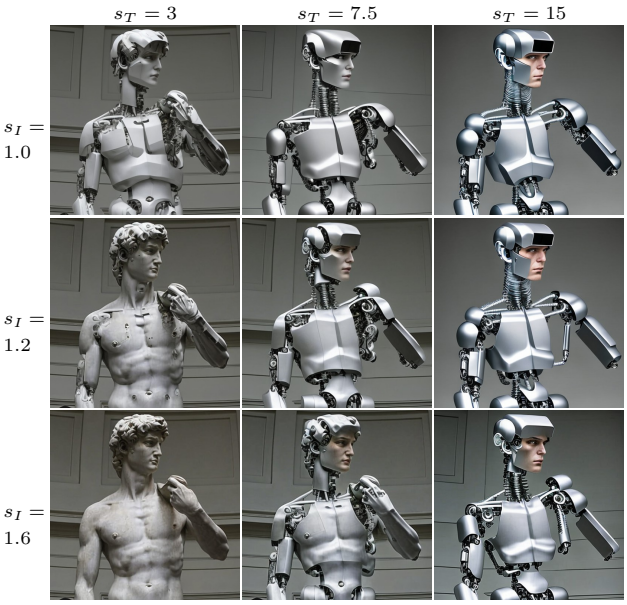


图 4. Classifier-free guidance weights over two conditional inputs. s_I controls similarity with the input image, while s_T controls consistency with the edit instruction. 两个条件输入上的无分类器引导权重。s_I 控制与输入图像的相似度，s_T 控制与编辑指令的一致性。

details are provided in the supplemental material.

Wang 等人 [67] 证明，在配对训练数据有限的情况下，对大型图像扩散模型进行微调优于从头训练图像翻译任务的模型。因此，我们使用预训练的 Stable Diffusion 检查点初始化模型权重，利用其强大的文本到图像生成能力。为支持图像条件，我们向第一卷积层添加额外的输入通道，连接 z_t 和 $\mathcal{E}(c_I)$。扩散模型的所有可用权重从预训练检查点初始化，对新添加输入通道操作的权重初始化为零。我们复用原本为描述文本设计的文本条件机制，改为接受文本编辑指令 c_T 作为输入。更多训练细节见附加材料。

3.2.1 面向两个条件的无分类器引导(Classifier-free Guidance for Two Conditionings)

Classifier-free diffusion guidance [20] is a method for trading off the quality and diversity of samples generated by a diffusion model. It is commonly used in class-conditional and text-conditional image generation to improve the visual quality of generated images and to make sampled images better correspond with their conditioning. Classifier-free guidance effectively shifts probability mass toward data where an implicit classifier $p_\theta(c|z_t)$ assigns high likelihood to the conditioning c. The implementation of classifier-free guidance involves jointly training the diffusion model for conditional and unconditional denoising, and combining the two score estimates at inference time. Training for unconditional denoising is done by simply setting the conditioning to a fixed null value $c{=}\varnothing$ at some frequency during training. At inference time, with a guidance scale $s \geq 1$, the modified score estimate $\tilde{e_\theta}(z_t, c)$ is extrapolated in the direction toward the conditional $e_\theta(z_t, c)$ and away from the unconditional $e_\theta(z_t, \varnothing)$.

无分类器扩散引导 [20] 是一种在扩散模型生成样本的质量和多样性之间进行权衡的方法。它通常用于类条件和文本条件图像生成，以改进生成图像的视觉质量，使采样图像更好地对应其条件。无分类器引导有效地将概率质量转移到隐式分类器 $p_\theta(c|z_t)$ 给条件 c 分配高似然的数据。无分类器引导的实现涉及联合训练扩散模型以进行条件和无条件去噪，并在推理时结合两个分数估计。通过在训练期间以某个频率简单地将条件设置为固定空值 $c{=}\varnothing$ 来进行无条件去噪训练。在推理时，使用引导尺度 $s \geq 1$，修改的分数估计 $\tilde{e_\theta}(z_t, c)$ 沿着指向条件 $e_\theta(z_t, c)$ 方向、远离无条件 $e_\theta(z_t, \varnothing)$ 方向进行外推。

$$\tilde{e_\theta}(z_t, c) = e_\theta(z_t, \varnothing) + s \cdot (e_\theta(z_t, c) - e_\theta(z_t, \varnothing)) \quad (2)$$

For our task, the score network $e_\theta(z_t, c_I, c_T)$ has two conditionings: the input image c_I and text instruction c_T. We find if beneficial to leverage classifier-free guidance with respect to both conditionings. Liu *et al.* [38] demonstrate that a conditional diffusion model can compose score estimates from multiple different conditioning values. We apply the same concept to our model with two separate conditioning inputs. During training, we randomly set only $c_I{=}\varnothing_I$ for 5% of examples, only $c_T{=}\varnothing_T$ for 5% of examples, and both $c_I{=}\varnothing_I$ and $c_T{=}\varnothing_T$ for 5% of examples. Our model is therefore capable of conditional or unconditional denoising with respect to both or either conditional inputs. We introduce two guidance scales, s_I and s_T, which can be adjusted to trade off how strongly the generated samples correspond with the input image and how strongly they correspond with the edit instruction. Our modified score estimate is as follows:

对于我们的任务，分数网络 $e_\theta(z_t, c_I, c_T)$ 有两个条件：输入图像 c_I 和文本指令 c_T。我们发现利用相对于

两个条件的无分类器引导是有益的。Liu 等人 [38] 证明条件扩散模型可以从多个不同条件值组合分数估计。我们对模型应用相同的概念，具有两个独立的条件输入。在训练期间，我们随机仅为 5% 的示例设置 $c_I = \varnothing_I$，仅为 5% 的示例设置 $c_T = \varnothing_T$，为 5% 的示例同时设置 $c_I = \varnothing_I$ 和 $c_T = \varnothing_T$。因此，我们的模型能够对所有、其中一个或两个条件输入进行条件或无条件去噪。我们引入两个引导尺度 s_I 和 s_T，可调整以权衡生成样本与输入图像的对应强度和与编辑指令的对应强度。我们修改的分数估计如下：

$$\begin{aligned}\tilde{e}_\theta(z_t, c_I, c_T) = e_\theta(z_t, \varnothing, \varnothing) \\ + s_I \cdot (e_\theta(z_t, c_I, \varnothing) - e_\theta(z_t, \varnothing, \varnothing)) \\ + s_T \cdot (e_\theta(z_t, c_I, c_T) - e_\theta(z_t, c_I, \varnothing))\end{aligned} \tag{3}$$

In Figure 4, we show the effects of these two parameters on generated samples. See Appendix B for details of our classifier-free guidance formulation.

在图 4 中，我们展示了这两个参数对生成样本的影响。详见附录 B 了解我们的无分类器引导公式的详细信息。

4. 结果（Results）

We show instruction-based image editing results on a diverse set of real photographs and artwork, for a variety of types of edits and instruction wordings. See Figures 1, 5, 6, 7, 11, 12, 15, 16, 17, 18, and 19 for selected results. Our model successfully performs many challenging edits, including replacing objects, changing seasons and weather, replacing backgrounds, modifying material attributes, converting artistic medium, and a variety of others.

我们展示了在真实照片与艺术作品的多样集合上展示指令指导的图像编辑成果。见图 1, 5, 6, 7, 11, 12, 15, 16, 17, 18, 和 19 的示例结果。我们的模型成功执行了多种具有挑战性的编辑，包括替换对象、改变季节和天气、替换背景、修改材质属性、转换艺术媒介等。

We compare our method qualitatively with a couple recent works, SDEdit [39] and Text2Live [6]. Our model follows instructions for how to edit the image, but prior works (including these baseline methods) expect descriptions of the image (or edit layer). Therefore, we provide them with the "after-edit" text caption instead of the edit instruction.

我们定性地将方法与两项最近的工作 SDEdit [39] 和 Text2Live [6] 进行比较。我们的模型遵循编辑图像的指令，而先前的方法（包括这些基线方法）需要图像的描述（或编辑图层）。因此，我们向它们提供编辑后的文本标题，而不是编辑指令。We also compare our method quantitatively with SDEdit, using two metrics measuring image consistency and edit quality, further described in Section 4.1. Finally, we show ablations on how the size and quality of generated training data affect our model's performance in Section 4.2.

我们还定量地将方法与 SDEdit 进行比较，使用两个衡量图像一致性和编辑质量的指标，在第 4.1 节中进一步说明。最后，我们在第 4.2 节展示了生成的训练数据的大小和质量如何影响模型性能的消融实验。

4.1. 基线对比（Baseline comparisons）

We provide qualitative comparisons with SDEdit [39] and Text2Live [6], as well as quantitative comparisons with SDEdit.

我们提供了与 SDEdit [39] 和 Text2Live [6] 的定性比较，以及与 SDEdit 的定量比较。SDEdit [39] is a technique for editing images with a pretrained diffusion model, where a partially noised image is passed as input and denoised to produce a new edited image. We compare with the public Stable Diffusion implementation of SDEdit. Text2Live [6] is a technique for editing images by generating a color+opacity augmentation layer, conditioned on a text prompt. We compare with the public implementation released by the authors.

SDEdit [39] 是一种使用预训练扩散模型编辑图像的技术，其中将部分加噪的图像作为输入并去噪以生成新的编辑图像。我们与 SDEdit 的公开 Stable Diffusion 实现进行比较。Text2Live [6] 是一种通过生成颜色 + 不透明度增强图层来编辑图像的技术，条件是文本提示。我们与作者公开发布的实现进行比较。

We compare with both methods qualitatively in Figure 9. We notice that while SDEdit works reasonably well for cases where content remains approximately constant and style is changed, it struggles to preserve identity and isolate individual objects, especially when larger changes are desired. Additionally, it requires a full output description of the desired image, rather than an editing instruction. On the other hand, while Text2Live is able to produce

图 5. *Mona Lisa* transformed into various artistic mediums. *Mona Lisa* 在多种艺术风格中的变换。

图 6. *The Creation of Adam* with new context and subjects (generated at 768 resolution). *The Creation of Adam* 的新背景和主题（768 分辨率生成）。

convincing results for edits involving additive layers, its formulation limits the categories of edits that it can handle.

我们在图 9 中定性地与两种方法进行比较。我们注意到，SDEdit 在内容保持近似恒定且风格改变的情况下效果相当好，但在保持身份和隔离单个对象方面存在困难，特别是当需要较大改变时。此外，它需要所需图像的完整输出描述，而不是编辑指令。另一方面，虽然 Text2Live 能够为涉及加法图层的编辑产生令人信服的结果，但其配置限制了它能处理的编辑类别。

Quantitative comparisons with SDEdit are shown in Figure 8. We plot the tradeoff between two metrics, cosine similarity of CLIP image embeddings (how much the edited image agrees with the input image) and the directional CLIP similarity introduced by [14] (how much the change in text captions agrees with the change in the images). These are competing metrics—increasing the degree to which the output images correspond to a desired edit will reduce their similarity (consistency) with the input image. Still, we find that when comparing our method with SDEdit, our results have notably higher image consistency (CLIP image similarity) for the same directional similarity values.

与 SDEdit 的定量比较如图 8 所示。我们绘制了两个指标之间的权衡：CLIP 图像嵌入的余弦相似度（编辑后图像与输入图像的一致程度）和 [14] 引入的方向性 CLIP 相似度（文本标题的变化与图像变化的一致程度）。这些是竞争指标——增加输出图像与所需编辑的对应程度会降低它们与输入图像的相似性（一致性）。尽管如此，我们发现在将方法与 SDEdit 进行比较时，在相同的方向性相似度值下，我们的结果具有明显更高的图像一致性（CLIP 图像相似度）。

4.2. 消融实验（Ablations）

In Figure 10, we provide quantitative ablations for both our choice of dataset size and our dataset filtering approach described in Section 3.1. We find that decreasing the size of the dataset typically results in decreased ability to perform larger (i.e., more significant) image edits, instead only performing subtle or stylistic image adjustments (and thus, maintaining a high image similarity score, but a low directional score). Removing the CLIP filtering from our dataset generation has a different effect: the overall image consistency with the input image is reduced.

在图 10 中，我们提供了数据集大小选择和数据集过滤方法（在第 3.1 节中描述）的定量消融。我们发现减小数据集大小通常会降低执行较大（即更显著）图像编辑的能力，而仅执行细微或风格化的图像调整（这样虽然保持高图像相似度，但方向相似度低）。从数据集生成中删除 CLIP 过滤有不同的效果：整体图像与输入图像的一致性会降低。

We also provide an analysis of the effect of our two classifier-free guidance scales in Figure 4. Increasing s_T results in a stronger edit applied to the image (i.e., the output agrees more with the instruction), and increasing s_I can help preserve the spatial structure of the input image (i.e., the output agrees more with the input image). We

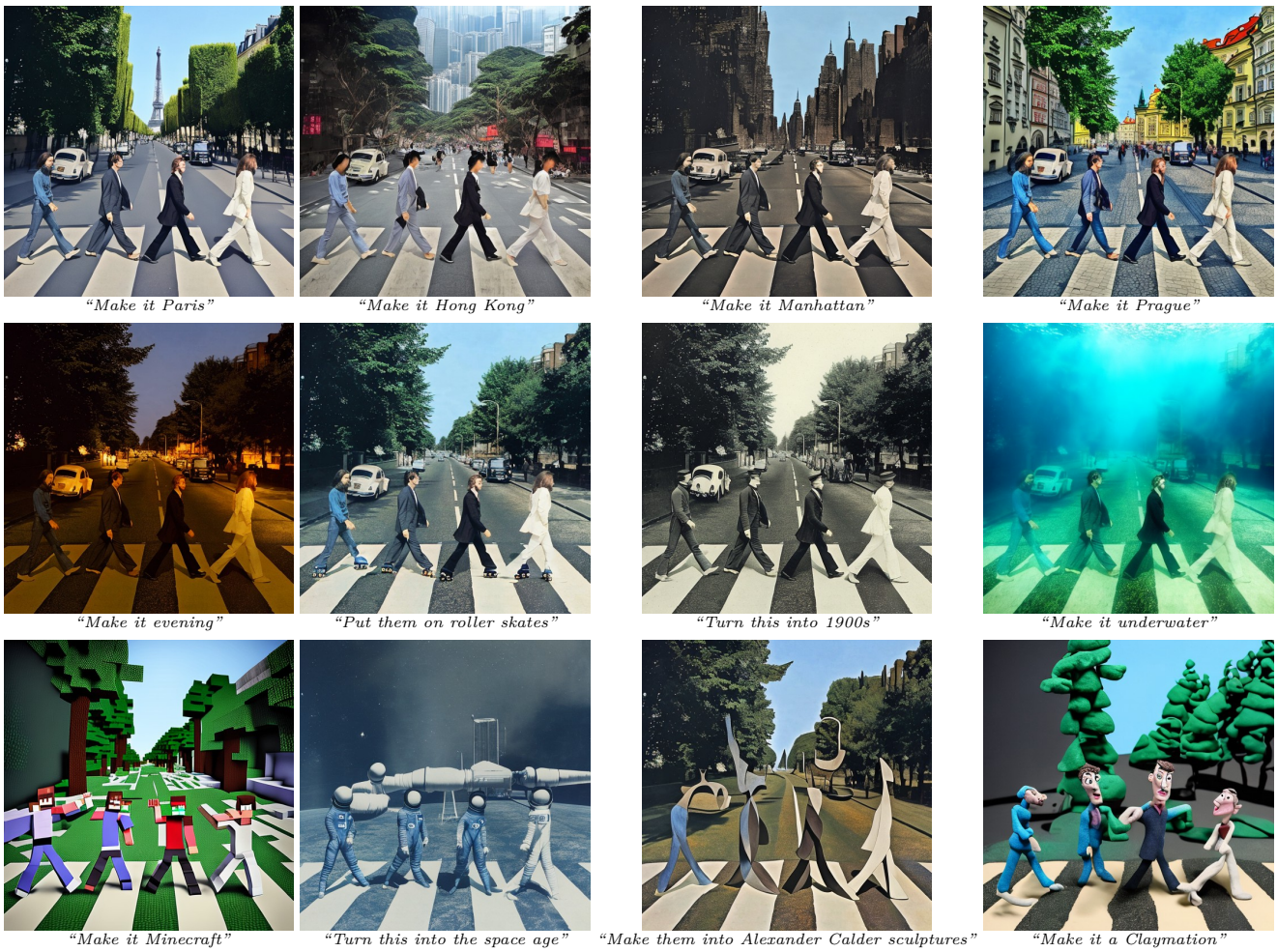


图 7. The iconic Beatles *Abbey Road* album cover transformed in a variety of ways. 披头士标志性专辑 *Abbey Road* 的各种变换。

find that values of s_T in the range 5−10 and values of s_I in the range 1−1.5 typically produce the best results. In practice, and for the results shown in the paper, we find it beneficial to adjust guidance weights for each example to get the best balance between consistency and edit strength.

我们还在图 4 中提供了两个无分类器引导尺度效果的分析。增加 s_T 会对图像进行更强的编辑（即输出更符合指令），增加 s_I 可以帮助保持输入图像的空间结构（即输出更符合输入图像）。我们发现 s_T 在 5−10 范围内的值和 s_I 在 1−1.5 范围内的值通常产生最佳结果。在实践中，对于论文中所示的结果，我们发现对每个示例调整引导权重以获得一致性和编辑强度之间的最佳平衡是有益的。

5. 讨论（Discussion）

We demonstrate an approach that combines two large pretrained models, a large language model and a text-to-image model, to generate a dataset for training a diffusion model to follow written image editing instructions. While our method is able to produce a wide variety of compelling edits to images, including style, medium, and other contextual changes, there still remain a number of limitations.

我们展示了一种方法，将两个大型预训练模型——语言模型和文本到图像模型——结合起来，生成数据集以训练扩散模型遵循书面图像编辑指令。虽然我们的方法能够对图像进行广泛的令人信服的编辑，包括风格、媒介和其他上下文改变，但仍存在许多局限性。

Our model is limited by the visual quality of the generated dataset, and therefore by the diffusion model used to

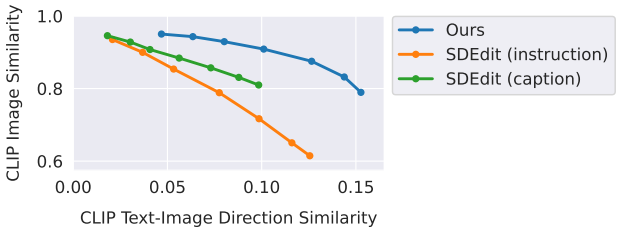


图 8. We plot the trade-off between consistency with the input image (Y-axis) and consistency with the edit (X-axis). For both metrics, higher is better. For both methods, we fix text guidance to 7.5, and vary our $s_I \in [1.0, 2.2]$ and SDEdit's strength (the amount of denoising) between $[0.3, 0.9]$. 我们绘制了与输入图像一致性（Y 轴）和与编辑一致性（X 轴）之间的权衡。对于两个指标，越高越好。对于两种方法，我们将文本引导固定为 7.5，改变 $s_I \in [1.0, 2.2]$ 和 SDEdit 的强度（去噪量）在 $[0.3, 0.9]$ 之间。

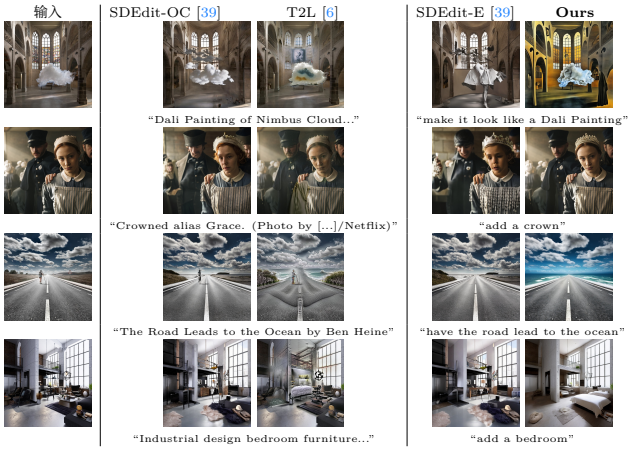


图 9. Comparison with other editing methods. The input is transformed either by edit string (last two columns) or the ground-truth output image caption (middle two columns). We compare our method against two recent works, SDEdit [39] and Text2Live [6]. We show SDEdit in two configurations: conditioned on the output caption (OP) and conditioned on the edit string (E). 与其他编辑方法的对比。输入通过编辑字符串（最后两列）或真值输出图像标题（中间两列）进行变换。我们将方法与两项最近的工作 SDEdit [39] 和 Text2Live [6] 进行比较。我们展示了两种 SDEdit 配置：以输出标题为条件（OC）和以编辑字符串为条件（E）。

generate the imagery (in this case, Stable Diffusion [52]). Furthermore, our method's ability to generalize to new edits and make correct associations between visual changes and text instructions is limited by the human-written instructions used to fine-tune GPT-3 [7], by the ability of GPT-3 to create instructions and modify captions, and by the ability of Prompt-to-Prompt [17] to modify generated images. In particular, our model struggles with counting numbers of objects and with spatial reasoning (e.g., *"move it to the left of the image"*, *"swap their positions"*, or *"put two cups on the table and one on the chair"*), just as in Stable Diffusion and Prompt-to-Prompt. Examples of failures can be found in Figure 13. Furthermore, there are well-documented biases in the data and the pretrained models that our method is based upon, and therefore the edited images from our method may inherit these biases or introduce other biases (Figure 14).

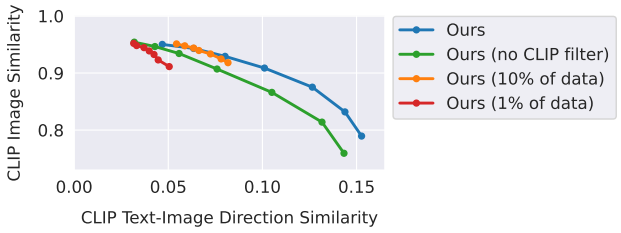
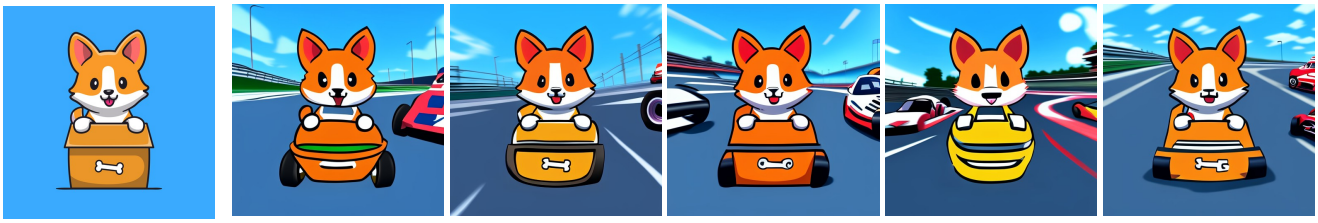


图 10. We compare ablated variants of our model (smaller training dataset, no CLIP filtering) by fixing s_T and sweeping values of $s_I \in [1.0, 2.2]$. Our proposed configuration performs best. 我们通过固定 s_T 并改变 $s_I \in [1.0, 2.2]$ 的值来比较模型的消融变体（较小的训练数据集、无 CLIP 过滤）。我们提议的配置性能最佳。

模型受到生成数据集视觉质量的限制，因此受到用于生成图像的扩散模型（在本例中为 Stable Diffusion [52]）的限制。此外，方法推广到新编辑并在视觉变化和文本指令之间进行正确关联的能力受限于用于微调 GPT-3 [7] 的人工编写指令、GPT-3 创建指令和修改标题的能力，以及 Prompt-to-Prompt [17] 修改生成图像的能力。特别是，模型在计数对象数量和空间推理方面存在困难（例如，*"move it to the left of the image"*、*"swap their positions"* 或 *"put two cups on the table and one on the chair"*），就像在 Stable Diffusion 和 Prompt-to-Prompt 中一样。故障示例可在图 13 中找到。此外，方法所基于的数据和预训练模型中存在文档记录的偏见，因此方法生成的编辑图像可能继承这些偏见或引入其他偏见（图 14）。

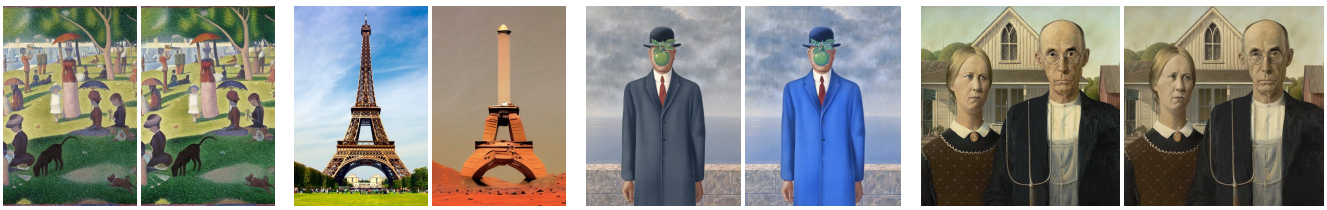
Aside from mitigating the above limitations, our work also opens up questions, such as: how to follow instructions for spatial reasoning, how to combine instructions with other conditioning modalities like user interaction, and

图 11. Applying our model recurrently with different instructions results in compounded edits. 使用不同指令递归应用模型会产生叠加编辑。



输入　　编辑指令: *"in a race car video game"*

图 12. By varying the latent noise, our model can produce many possible image edits for the same input image and instruction. 通过改变潜在噪声，模型可以为相同的输入图像和指令生成许多可能的图像编辑。



"Zoom into the image"　*"Move it to Mars"*　*"Color the tie blue"*　*"Have the people swap places"*

图 13. Failure cases. Left to right: our model is not capable of performing viewpoint changes, can make undesired excessive changes to the image, can sometimes fail to isolate the specified object, and has difficulty reorganizing or swapping objects with each other. 失败案例。从左到右：模型无法执行视角变化，可能对图像进行不希望的过度改变，有时无法隔离指定对象，并且在重新组织或交换对象方面存在困难。

how to evaluate instruction-based image editing. Incorporating human feedback to improve the model is another important area of future work, and strategies like human-in-the-loop reinforcement learning could be applied to improve alignment between our model and human intentions.

除了缓解上述局限性外，我们的工作还提出了问题，如：如何遵循空间推理的指令、如何将指令与其他条件模态（如用户交互）结合，以及如何评估指令指导的图像编辑。结合人类反馈来改进模型是未来工作的另一个重要领域，可以应用人类在环强化学习等策略来改进模型与人类意图之间的对齐。

致 谢 （Acknowledgments） We thank Ilija Radosavovic, William Peebles, Allan Jabri, Dave Epstein, Kfir Aberman, Amanda Buster, and David Salesin. Tim Brooks is funded by an NSF Graduate Research Fellowship. Additional funding provided by a research grant from SAP and a gift from Google.

参考文献

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4432–4441, 2019. 3

输入 *"Make them look like flight attendants"* *"Make them look like doctors"*

图 14. Our method reflects biases from the data and models it is based upon, such as correlations between profession and gender. 我们的方法反映了其所基于的数据和模型中的偏见，例如职业与性别之间的相关性。

输入 *"Make it a grocery store"*

图 15. Leighton's *Lady in a Garden* moved to a new setting. Leighton 的 *Lady in a Garden* 置于新环境。

[2] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8296–8305, 2020. 3

[3] Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit Bermano. Hyperstyle: Stylegan inversion with hypernetworks for real image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18511–18521, 2022. 3

[4] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 2

[5] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022. 3, 4

[6] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *European Conference on Computer Vision*, pages 707–723. Springer, 2022. 3, 4, 9, 12

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing sys-*

输入 *"Convert to a realistic photo"* *"Turn into a 3D model"*

图 16. Van Gogh's *Self-Portrait with a Straw Hat* in different mediums. Van Gogh 的 *Self-Portrait with a Straw Hat* 以不同的媒介呈现。

输入 *"Add boats on the water"* *"Replace the mountains with a city skyline"*

图 17. A landscape photograph shown with different contextual edits. Note that isolated changes also bring along accompanying contextual effects: the addition of boats also adds wind ripples in the water, and the added city skyline is reflected on the lake. 风景照片展示了不同的上下文编辑。注意隔离的变化也会带来伴随的上下文效果：添加船只也会在水中增加风波纹，添加的城市天际线也会在湖面上反射。

输入 *"It is now midnight"* *"Add a beautiful sunset"*

图 18. A photograph of a cityscape edited to show different times of day. 展示不同时间的城市景观照片编辑。

tems, 33:1877–1901, 2020. 2, 4, 5, 12

[8] Lucy Chai, Jonas Wulff, and Phillip Isola. Using latent space regression to analyze and leverage compositionality in gans. In *International Conference on Learning Representations*, 2021. 3

[9] Katherine Crowson, Stella Biderman, Daniel Kornis,

图 19. Vermeer's *Girl with a Pearl Earring* with a variety of edits. Vermeer 的 *Girl with a Pearl Earring* 的多种编辑。

Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance. In *European Conference on Computer Vision*, pages 88–105. Springer, 2022. 3

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 2

[11] Yilun Du, Shuang Li, and Igor Mordatch. Compositional visual generation with energy based models. *Advances in Neural Information Processing Systems*, 33:6637–6647, 2020. 2

[12] Dave Epstein, Taesung Park, Richard Zhang, Eli Shechtman, and Alexei A Efros. Blobgan: Spatially disentangled scene representations. In *European Conference on Computer Vision*, pages 616–635. Springer, 2022. 3

[13] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 3, 4

[14] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022. 3, 7, 10

[15] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015. 3

[16] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 3

[17] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 3, 4, 6, 12, 20

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural*

Information Processing Systems, 33:6840–6851, 2020. 2

[19] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 23:47–1, 2022. 2

[20] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 8

[21] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022. 2

[22] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018. 3

[23] Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005. 7, 21

[24] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017. 3

[25] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *arXiv preprint arXiv:2206.00364*, 2022. 20, 21

[26] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 3

[27] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 3

[28] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*, 2022. 3, 4

[29] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435, 2022. 3

[30] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 7

[31] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020. 2

[32] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18062–18071, 2022. 3

[33] Daiqing Li, Huan Ling, Seung Wook Kim, Karsten Kreis, Sanja Fidler, and Antonio Torralba. Bigdatasetgan: Synthesizing imagenet with pixel-wise annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21330–21340, 2022. 4

[34] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 2

[35] Shuang Li, Yilun Du, Joshua B Tenenbaum, Antonio Torralba, and Igor Mordatch. Composing ensembles of pre-trained models via iterative consensus. *arXiv preprint arXiv:2210.11522*, 2022. 2

[36] Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *arXiv preprint arXiv:2205.14217*, 2022. 2

[37] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 3

[38] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. *arXiv preprint arXiv:2206.01714*, 2022. 2, 8, 9

[39] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2021. 3, 4, 9, 12

[40] Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. *arXiv preprint arXiv:2104.08773*, 2021. 4

[41] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 2

[42] 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. *arXiv preprint arXiv:2112.10741*, 2021. 2, 3

[43] Utkarsh Ojha, Yijun Li, Cynthia Lu, Alexei A. Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. Few-shot image generation via cross-domain correspondence. In *CVPR*, 2021. 3

[44] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022. 4

[45] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2085–2094, October 2021. 3

[46] William Peebles, Ilija Radosavovic, Tim Brooks, Alexei A Efros, and Jitendra Malik. Learning to learn with generative models of neural network checkpoints. *arXiv preprint arXiv:2209.12892*, 2022. 2

[47] William Peebles, Jun-Yan Zhu, Richard Zhang, Antonio Torralba, Alexei A Efros, and Eli Shechtman. Gan-supervised dense visual alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13470–13481, 2022. 4

[48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 3

[49] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2, 3, 4

[50] Suman Ravuri and Oriol Vinyals. Classification accuracy score for conditional generative models. *Advances in neural information processing systems*, 32, 2019. 4

[51] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2287–2296, 2021. 3

[52] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2, 3, 4, 5, 6, 7, 12, 20

[53] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022. 3, 4

[54] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 2

[55] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 2

[56] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[57] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022. 5

[58] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2107–2116, 2017. 4

[59] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2

[60] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. 2, 7

[61] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[62] Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. Zero-shot image-to-text generation for visual-semantic arithmetic. *arXiv preprint arXiv:2111.14447*, 2021. 2

[63] Anthony Meng Huat Tiong, Junnan Li, Boyang Li, Silvio Savarese, and Steven CH Hoi. Plug-and-play vqa: Zero-shot vqa by conjoining large pretrained models with zero training. *arXiv preprint arXiv:2210.08773*, 2022. 2

[64] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021. 3

[65] Nontawat Tritrong, Pitchaporn Rewatbowornwong, and Supasorn Suwajanakorn. Repurposing gans for one-shot semantic part segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4475–4485, 2021. 4

[66] Yuri Viazovetskyi, Vladimir Ivashkin, and Evgeny Kashin. Stylegan2 distillation for feed-forward image manipulation. In *European conference on computer vision*, pages 170–186. Springer, 2020. 4

[67] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation. *arXiv preprint arXiv:2205.12952*, 2022. 7, 8

[68] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 2

[69] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021. 4

[70] Andy Zeng, Adrian Wong, Stefan Welker, Krzysztof Choromanski, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, et al. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*, 2022. 2

[71] Wanfeng Zheng, Qiang Li, Xiaoyan Guo, Pengfei Wan, and Zhongyuan Wang. Bridging clip and stylegan through latent alignment for image editing. *arXiv preprint arXiv:2210.04506*, 2022. 3

[72] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017. 3

A. 实现细节（Implementation Details）

A.1. 指令与描述文本生成（Instruction and Caption Generation）

We finetune GPT3 to generate edit instructions and edited captions. The text prompt used during fine-tuning is the input caption concatenated with `"\n##\n"` as a separator token. The text completion is a concatenation of the instruction and edited caption with `"\n%%\n"` as a separator token in between the two and `"\nEND"` appended to the end as the stop token. During inference, we sample text completions given new input captions using `temperature=0.7` and `frequency_penalty=0.1`. We exclude generations where the input and output captions are the same.

我们对 GPT3 进行微调以生成编辑指令和编辑后的描述文本。微调期间使用的文本提示是输入描述文本，后跟 `"\n##\n"` 作为分隔符 token。文本完成是指令和编辑后描述文本的连接，中间以 `"\n%%\n"` 作为分隔符 token，末尾追加 `"\nEND"` 作为停止符 token。推理时，对新的输入描述文本采样文本完成，使用 `temperature=0.7` 和 `frequency_penalty=0.1`。我们排除输入和输出描述文本相同的生成结果。

A.2. 配对图像生成（Paired Image Generation）

We generate paired before/after training images from paired before/after captions using Stable Diffusion [52] in combination with Prompt-to-Prompt [17]. We use exponential moving average (EMA) weights of the Stable Diffusion v1.5 checkpoint and the improved ft-MSE autoencoder weights. We generate images with 100 denoising steps using an Euler ancestral sampler with denoising variance schedule proposed by Kerras *et al.* [25]. We ensure the same latent noise is used for both images in each generated pair (for initial noise as well as noise introduced during stochastic sampling).

我们使用 Stable Diffusion [52] 结合 Prompt-to-Prompt [17]，从配对的编辑前后描述文本生成配对的编辑前后训练图像。我们使用 Stable Diffusion v1.5 检查点的指数移动平均（EMA）权重和改进的 ft-MSE 自编码器权重。我们使用 Euler ancestral 采样器生成图像，共进行 100 个去噪步，去噪方差调度由 Kerras *et al.* [25] 提出。我们确保每个生成的配对中两张图像使用相同的潜在噪声（包括初始噪声和随机采样期间引入的噪声）。

Prompt-to-Prompt replaces cross-attention weights in the second generated image differently based on the specific edit type: word swap, adding a phrase, increasing or decreasing weight of a word. We instead replaced *self*-attention weights of the second image for the first p fraction of steps, and use the same attention weight replacement strategy for all edits.

Prompt-to-Prompt 根据具体的编辑类型不同地替换第二张生成图像中的交叉注意力权重：词替换、添加短语、增加或减少词的权重。我们改为在前 p 比例的步数中替换第二张图像的自注意力权重，并对所有编辑使用相同的注意力权重替换策略。

We generation 100 pairs of images for each pair of captions. We filter training data for an image-image CLIP threshold of 0.75 to ensure images are not too different, an image-caption CLIP threshold of 0.2 to ensure images correspond with their captions, and a directional CLIP similarity of 0.2 to ensure the change in before/after captions correspond with the change in before/after images. For each each pair of captions, we sort any image pairs that pass all filters by the directional CLIP similarity and keep up to 4 examples.

我们为每对描述文本生成 100 个图像对。我们使用图像-图像 CLIP 阈值 0.75 对训练数据进行过滤，以确保图像差异不过大；图像-描述 CLIP 阈值 0.2，以确保图像与其描述文本对应；方向性 CLIP 相似度 0.2，以确保编辑前后描述文本的变化与编辑前后图像的变化相对应。对于每对描述文本，我们按方向性 CLIP 相似度排序通过所有过滤器的图像对，并保留最多 4 个示例。

A.3. 训练 InstructPix2Pix（Training InstructPix2Pix）

We train our image editing model for 10,000 steps on 8× 40GB NVIDIA A100 GPUs over 25.5 hours. We train at 256×256 resolution with a total batch size of 1024. We apply random horizontal flip augmentation and crop augmentation where images are first resized randomly between 256 and 288 pixels and then cropped to 256. We use a learning rate of 10^{-4} (without any learning rate warm up). We initialize our model from EMA weights of the Stable Diffusion v1.5 checkpoint, and adopt other training settings from the public Stable Diffusion code base.

我们在 8 张 40GB NVIDIA A100 GPU 上训练图

像编辑模型，共 10000 步，耗时 25.5 小时。我们以 256×256 分辨率训练，总批大小为 1024。我们应用随机水平翻转增强和裁剪增强，其中图像首先被随机调整为 256 到 288 像素之间，然后裁剪到 256。我们使用学习率 10^{-4}（无学习率预热）。我们从 Stable Diffusion v1.5 检查点的 EMA 权重初始化模型，并采用公开 Stable Diffusion 代码库中的其他训练设置。

While our model is trained at 256×256 resolution, we find it generalized well to 512×512 resolution at inference time, and generate results in this paper at 512 resolution with 100 denoising steps using an Euler ancestral sampler with denoising variance schedule proposed by Kerras *et al.* [25]. Editing an image with our model takes roughly 9 seconds on an A100 GPU.

虽然我们的模型以 256×256 分辨率训练，但我们发现它在推理时能很好地泛化到 512×512 分辨率。本文以 512 分辨率生成结果，使用 100 个去噪步，采用 Euler ancestral 采样器和由 Kerras *et al.* [25] 提出的去噪方差调度。在 A100 GPU 上编辑一张图像大约需要 9 秒。

B. 无分类器引导细节（Classifier-free Guidance Details）

As discussed in Section 3.2.1, we apply classifier-free guidance with respect to two conditionings: the input image c_I and the text instruction c_T. We introduce separate guidance scales s_I and s_T that enable separately trading off the strength of each conditioning. Below is the modified score estimate for our model with classifier-free guidance (copied from Equation 3):

如第 3.2.1 节所讨论的，我们相对于两个条件进行无分类器引导：输入图像 c_I 和文本指令 c_T。我们引入单独的引导尺度 s_I 和 s_T，使得能够分别权衡每个条件的强度。以下是我们模型应用无分类器引导的修改后的分数估计（复制自方程 3）：

$$\begin{aligned}\tilde{e_\theta}(z_t, c_I, c_T) = e_\theta(z_t, \varnothing, \varnothing) \\ + s_I \cdot (e_\theta(z_t, c_I, \varnothing) - e_\theta(z_t, \varnothing, \varnothing)) \\ + s_T \cdot (e_\theta(z_t, c_I, c_T) - e_\theta(z_t, c_I, \varnothing))\end{aligned}$$

Our generative model learns $P(z|c_I, c_T)$, the probability distribution of image latents $z = \mathcal{E}(x)$ conditioned on an input image c_I and a text instruction c_T. We arrive at our particular classifier-free guidance formulation by expressing the conditional probability as follows:

我们的生成模型学习 $P(z|c_I, c_T)$，即以输入图像 c_I 和文本指令 c_T 为条件的图像潜在变量 $z = \mathcal{E}(x)$ 的概率分布。我们通过如下方式表达条件概率来得到特定的无分类器引导公式：

$$P(z|c_T, c_I) = \frac{P(z, c_T, c_I)}{P(c_T, c_I)} = \frac{P(c_T|c_I, z)P(c_I|z)P(z)}{P(c_T, c_I)}$$

Diffusion models estimate the score [23] of the data distribution, i.e., the derivative of the log probability. Taking the logarithm gives us the following expression:

扩散模型估计数据分布的分数 [23]，即对数概率的导数。取对数得到如下表达式：

$$\begin{aligned}\log(P(z|c_T, c_I)) = \log(P(c_T|c_I, z)) + \log(P(c_I|z)) \\ + \log(P(z)) - \log(P(c_T, c_I))\end{aligned}$$

Taking the derivative and rearranging we attain:

对导数进行求解和重新排列，我们得到：

$$\begin{aligned}\nabla_z \log(P(z|c_T, c_I)) = \nabla_z \log(P(z)) \\ + \nabla_z \log(P(c_I|z)) \\ + \nabla_z \log(P(c_T|c_I, z))\end{aligned}$$

This corresponds with the terms in our classifier-free guidance formulation in Equation 3. Our guidance scale s_I effectively shifts probability mass toward data where an implicit classifier $p_\theta(c_I|z_t)$ assigns high likelihood to the image conditioning c_I, and our guidance scale s_T effectively shifts probability mass toward data where an implicit classifier $p_\theta(c_T|c_I, z_t)$ assigns high likelihood to the text instruction conditioning c_T. Our model is capable of learning these implicit classifiers by taking the differences between estimates with and without the respective conditional input. Note there are multiple possible formulations such as switching the positions of c_T and c_I variables. We found that our particular decomposition works better for our use case in practice.

这与方程 3 中无分类器引导公式的项相对应。我们的引导尺度 s_I 有效地将概率质量转移到隐式分类器 $p_\theta(c_I|z_t)$ 为图像条件 c_I 分配高似然度的数据；我们

的引导尺度 s_T 有效地将概率质量转移到隐式分类器 $p_\theta(c_T|c_I, z_t)$ 为文本指令条件 c_T 分配高似然度的数据。我们的模型通过计算有无各自条件输入的估计之间的差异来学习这些隐式分类器。注意存在多种可能的公式表达，例如交换 c_T 和 c_I 变量的位置。我们发现在实践中，我们特定的分解对我们的用例效果更好。