西北工业大学

数字图像处理-论文翻译

原论文标题：InstructPix2Pix: Learning to Follow Image Editing Instructions

梁可怡

计算机学院

计算机科学与技术

2024 年 11 月

学号：2022304081

Northwestern Polytechnical University

InstructPix2Pix: Learning to Follow Image Editing Instructions

Tim Brooks* Aleksander Holynski* Alexei A. Efros
University of California, Berkeley

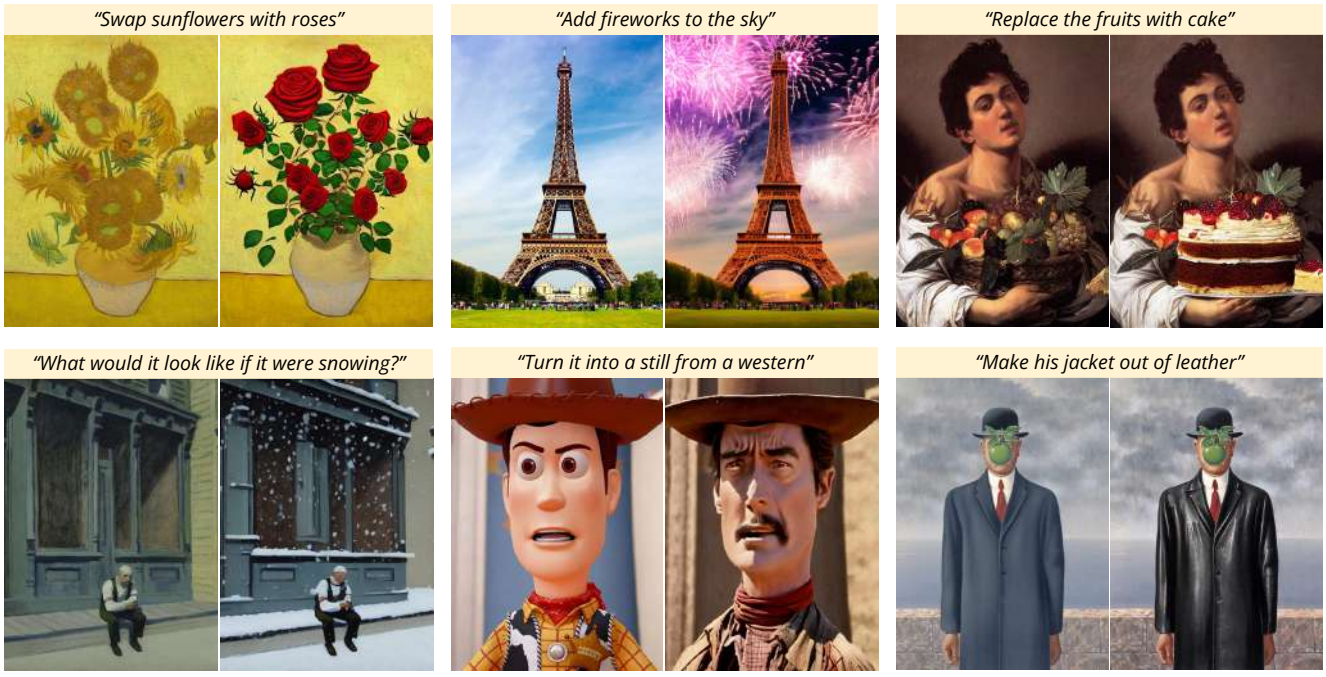

Figure 1. Given **an image** and **an instruction** for how to edit that image, our model performs the appropriate edit. Our model does not require full descriptions for the input or output image, and edits images in the forward pass without per-example inversion or fine-tuning.

Abstract

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.

*Denotes equal contribution
More results on our project page: timothybrooks.com/instruct-pix2pix

1. 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 [51]). 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.

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 fine-tuning. 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.

2. Prior work

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, 33, 40, 67], communication through prompting [62, 69], composing probability distributions of energy-based models [11, 37], guiding one model with feedback from another [61], and iterative optimization [34]. Our method is similar to prior work in that it leverages the complementary abilities of two pretrained models—GPT-3 [7]) and Stable Diffusion [51]—but differs in that we use these models to generate paired multi-modal training data.

Diffusion-based generative models Recent advances in diffusion models [59] have enabled state-of-the-art image synthesis [10, 18, 19, 53, 55, 60] as well as generative models of other modalities such as video [21, 58], audio [30], text [35] and network parameters [45]. Recent text-to-image diffusion models [41, 48, 51, 54] have shown to generate realistic images from arbitrary text captions.

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, 36, 42, 71]. Numerous editing approaches invert [1–3, 12] or encode [8, 50, 63] images into a latent space (e.g., StyleGAN [25, 26]) where they can be edited by manipulating latent vectors. Recent models have leveraged CLIP [47] embeddings to guide image editing using text [5, 9, 14, 28, 31, 41, 44, 70]. 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.

Recent works have used pretrained text-to-image diffusion models for image editing [5, 17, 27, 38, 48]. 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 [48]), 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 [38] 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, 48], generate new images of a specific object or concept learned from a small collection of images [13, 52], or perform editing by inverting (and fine-tuning) a single image, and subsequently regenerating with a new text description [27]. 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 finetuning.

Learning to follow instructions Our method differs from existing text-based image editing works [6, 13, 17, 27, 38, 52] 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 [39, 43, 68].

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 [32, 46, 49, 57, 64, 65]. In this paper, we use two different off-the-shelf generative models (language, text-to-image) to produce training data for our editing model.

3. 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

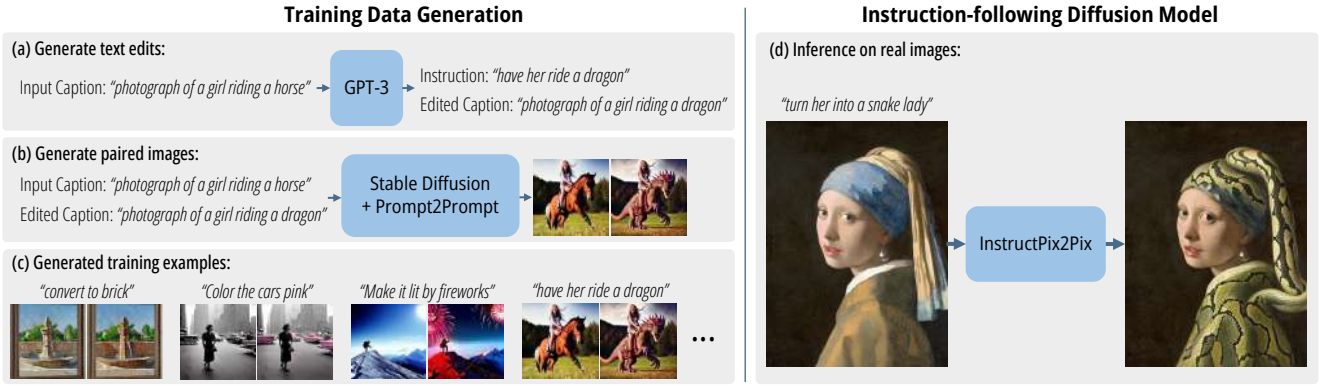

Figure 2. Our method consists of two parts: generating an image editing dataset, and training a diffusion model on that dataset. (a) We first use a finetuned GPT-3 to generate instructions and edited captions. (b) We then use StableDiffusion [51] in combination with Prompt-to-Prompt [17] to generate 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.

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.

3.1. Generating a Multi-modal 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 [51]—to generate 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 [51] (Figure 2b).

3.1.1 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.

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+ [56] 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.

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.

3.1.2 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

	Input LAION caption	Edit instruction	Edited caption
Human-written (700 edits)	*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 generated (>450,000 edits)	*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*
	...	...	...

Table 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.

(a) Without Prompt-to-Prompt. (b) With Prompt-to-Prompt.

Figure 3. Pair of images generated using StableDiffusion [51] 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"*.

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.

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.

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.

Diffusion models [59] 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 [51] improves the efficiency and quality of diffusion models by operating in the latent space of a pretrained variational autoencoder [29] 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:

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

Wang *et al.* [66] 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, lever-

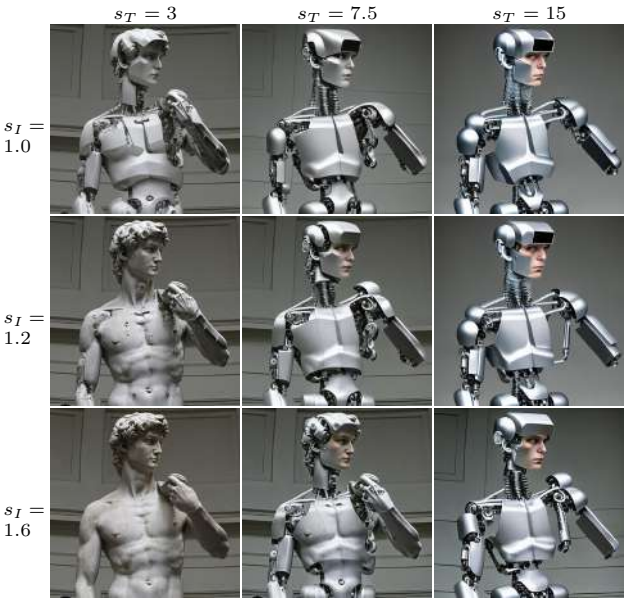

Figure 4. Classifier-free guidance weights over two conditional inputs. Higher values of s_I produce edited images with spatial structure more similar to the input image, and higher values of s_T produce images with more intense edits.

aging 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 details are provided in Appendix C of the supplement.

3.2.1 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 $e_\theta(z_t, \varnothing)$.

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

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.* [37] 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. Increasing s_I results in edited images that more closely resemble the input image, and increasing s_T results in more intense edits. Our modified score estimate is as follows:

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

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

4. Results

We show instruction-based image editing results on a diverse set of real photographs and artwork, for many edit types and instruction wordings. See Figures 1, 5, 6, 7, 11, 12 and Appendix A in the supplement 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.

We compare our method qualitatively with recent works SDEdit [38], Text2Live [6], and Prompt-to-Prompt [17]. 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. We also compare our method quantitatively with SDEdit and Prompt-to-Prompt, 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.

Figure 5. *Mona Lisa* transformed into various artistic mediums.

Figure 6. *The Creation of Adam* with new context and subjects (generated at 768 resolution).

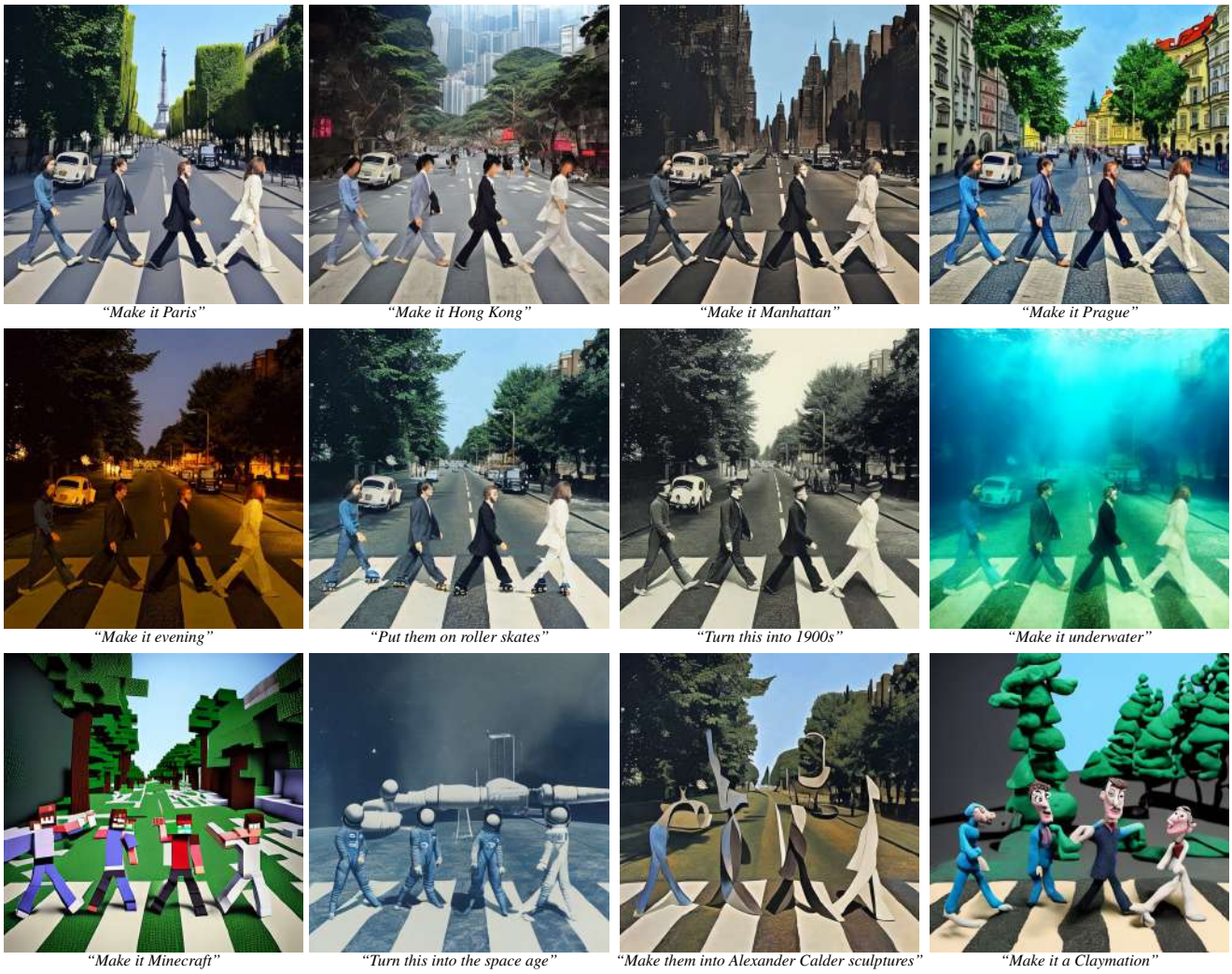

Figure 7. The iconic Beatles *Abbey Road* album cover transformed in a variety of ways.

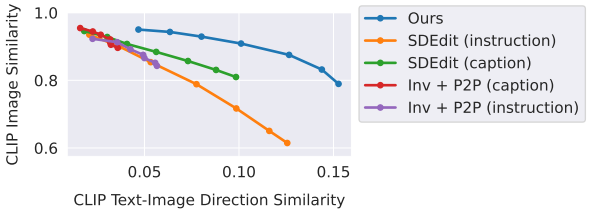

Figure 8. We plot the trade-off between consistency with the input image (Y-axis) and consistency with the edit (X-axis). For both metrics, higher is better. We fix text guidance to 7.5, and vary: our method's $s_I \in [1.0, 2.2]$, SDEdit's strength (the amount of denoising) in $[0.3, 0.9]$, and Prompt-to-Prompt's cross-attention period in $[0, 1]$. We experiment with two variants of Prompt-to-Prompt, using either the output caption or edit instruction.

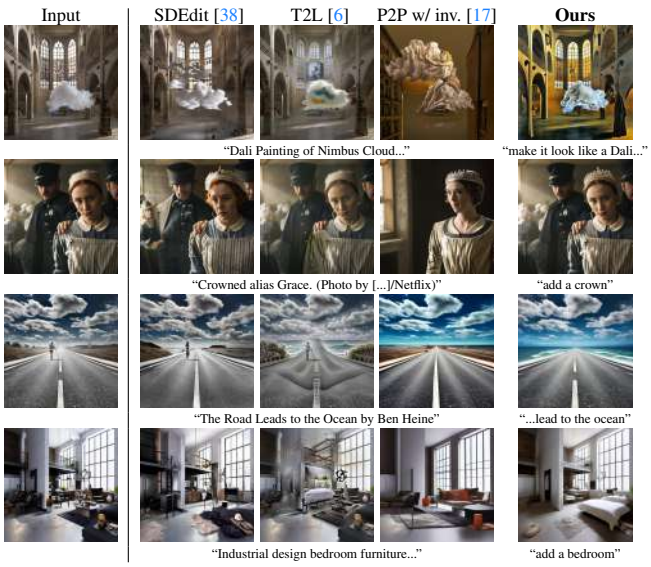

Figure 9. Qualitative comparison. We compare with recent image editing approaches SDEdit [38], Text2Live [6], and Prompt-to-Prompt [17]. These methods all expect an output image caption, unlike our method, which follows an editing instruction.

4.1. Baseline comparisons

We provide qualitative comparisons with SDEdit [38], Text2Live [6], and Prompt-to-Prompt [17], as well as quantitative comparisons with SDEdit and Prompt-to-Prompt. SDEdit [38] is a technique for editing images with a pre-trained diffusion model, where a partially noised image is passed as input and denoised to produce a new edited image. Text2Live [6] edits images by generating a color+opacity augmentation layer, conditioned on a text prompt.

We compare with SDEdit, Text2Live, and Prompt-to-Prompt qualitatively in Figure 9. Additional comparisons on other examples, as well as other configurations of these related works are provided in Appendix B of the supplement. 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

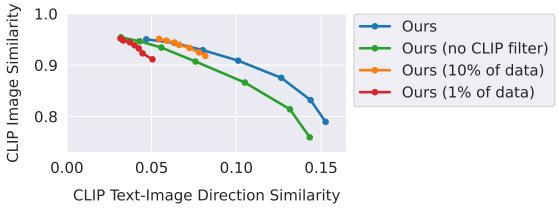

Figure 10. We compare ablated variants of our model (smaller training dataset, no CLIP filtering) by fixing s_T and sweeping values of $s_I \in [1.0, 2.2]$. Our proposed configuration performs best.

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

Quantitative comparisons with SDEdit and Prompt-to-Prompt 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 correspond to a desired edit will reduce its similarity with the input image—and we are interested in which method achieves the best tradeoff (highest curve). We find that compared to SDEdit and Prompt-to-Prompt, our results achieve higher directional similarity for the same image similarity values, indicating it better performs the desired edit. These findings are measured on average across 2000 edits, and we further validate them using a different CLIP model in Fig. 23 of the supplement. Outperforming Prompt-to-Prompt may seem surprising, since it is used in our training data generation, and in Appendix B we discuss possible causes for this improvement.

4.2. Ablations

In Fig. 10, we provide quantitative ablations for both our choice of dataset size and our dataset filtering approach described in Section 3.1. Decreasing the size of the dataset typically results in decreased ability to perform 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). In contrast, removing the CLIP filtering from our dataset generation reduces the overall image consistency with the input image.

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 find that

Figure 11. Applying our model recurrently with different instructions results in compounded edits.

Input

Edit instruction: *"in a race car video game"*

Figure 12. By varying the latent noise, our model can produce many possible image edits for the same input image and instruction.

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

Figure 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.

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.

5. 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.

Our model is limited by the visual quality of the generated dataset, and therefore by the diffusion model used to generate the imagery (in this case, Stable Diffusion [51]). 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. Additionally, we find that performing many sequential edits sometimes causes accumulating artifacts. 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. The edited images from our method may inherit these biases or introduce others (Fig. 14 in the supplement).

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 how to evaluate instruction-based editing. Incorporating human feedback, such as with the use of reinforcement learning, is another important direction for future work and could improve alignment between our model and human intentions.

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.

References

[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. 2

[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. 2

[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. 2

[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. 2

[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. 2, 5, 7

[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 systems*, 33:1877–1901, 2020. 1, 2, 3, 8

[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. 2

[9] Katherine Crowson, Stella Biderman, Daniel Kornis, 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. 2

[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. 2

[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. 2

[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. 2, 4, 7

[15] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015. 2

[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. 2

[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. 2, 3, 4, 5, 7, 8

[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. 5

[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. 2

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

[24] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017. 2

[25] 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. 2

[26] 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. 2

[27] 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. 2

[28] 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. 2

[29] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 4

[30] 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

[31] 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. 2

[32] 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. 2

[33] 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

[34] 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

[35] 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

[36] 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. 2

[37] 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, 5

[38] 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. 2, 5, 7

[39] Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. *arXiv preprint arXiv:2104.08773*, 2021. 2

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

[41] 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

[42] 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. 2

[43] 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. 2

[44] 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. 2

[45] 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

[46] 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. 2

[47] 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. 2

[48] 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

[49] Suman Ravuri and Oriol Vinyals. Classification accuracy score for conditional generative models. *Advances in neural information processing systems*, 32, 2019. 2

[50] 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. 2

[51] 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. 1, 2, 3, 4, 8

[52] 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. 2

[53] 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

[54] 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

[55] 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

[56] 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. 3

[57] 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. 2

[58] 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

[59] 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, 4

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

[61] 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

[62] 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

[63] 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. 2

[64] 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. 2

[65] 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. 2

[66] 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. 4

[67] 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

[68] 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. 2

[69] 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

[70] 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. 2

[71] 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. 2

InstructPix2Pix：学习遵循图像编辑指令

Tim Brooks*　Aleksander Holynski*　Alexei A. Efros

加州大学伯克利分校

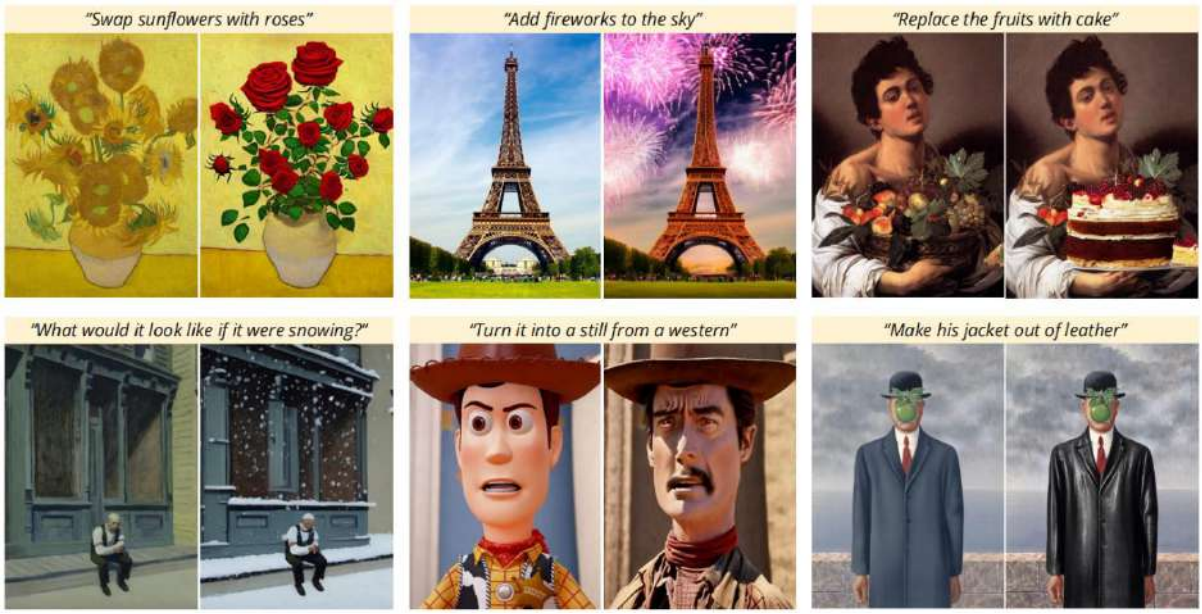

图 1. 如果给定一幅**图像**和如何编辑该图像的**指令**，我们的模型就会执行相应的编辑。我们的模型不需要输入或输出图像的完整描述，并且在前向传播中编辑图像，无需按实例反转或微调。

摘要

我们提出了一种根据人类指令编辑图像的方法：给定一张输入图像和一条书面指令，告诉模型应该做什么，我们的模型就按照这些指令编辑图像。为了获得这个任务的训练数据，我们结合了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（稳定扩散模型）——来生成一个大型图像编辑示例数据集。我们的条件扩散模型 InstructPix2Pix 是在我们生成的数据上训练出来的，在推理时可泛化为真实图像和用户编写的指令。由于该模型在前向传播中执行编辑，不需要对每个示例进行微调或反转，因此能在几秒钟内快速编辑图像。我们展示了针对各种输入图像和书面指令的令人信服的编辑结果。

1. 引言

我们介绍了一种教学方法，让生成模型按照人类编写的指令进行图像编辑。由于该任务的训练数据难以大规模获取，我们提出了一种生成配对数据集的方法，该数据集结合了在不同模式下预先训练的多个大模型：一个大语言模型（GPT-3[7]）和一个文本到图像模型（稳定扩散模型[51]）。这两个模型捕捉了语言和图像的互补知识，可以结合起来为跨越两种模态的任务创建配对训练数据。

利用我们生成的配对数据，我们训练了一个条件扩散模型，该模型在给定输入图像和如何编辑图像的文本指令后，生成编辑后的图像。我们的模型在前向传播中直接执行图像编辑，不需要任何额外的示例图像、输入/输出图像的完整描述或每个示例的微调。尽管我们的模型完全是在合成示例（即生成的书面指令和生成的图像）上训练出来的，但它对真实图像和自然的人类书面指令都实现了零误差泛化。我们的模型实现了直观的图像编辑，可以按照人类的指令进行各种编辑：替换对象、改变图像风格、改变场景和艺术媒介等。部分示例见图 1。

2. 先前的工作

组合大型预训练模型 最近的工作表明，可以将大型预训练模型组合起来，以解决任何一个模型都无法单独完成的多模态任务，例如图像描述和视觉问答（这些任务需要大语言模型和文本图像模型的知识）。组合预训练模型的技术包括在新任务中联合微调[4, 33, 40, 67]、通过提示进行交流[62, 69]、组合基于能量模型的概率分布[11, 37]、用另一个模型的反馈指导一个模型[61]以及迭代优化[34]。我们的方法与之前的工作类似，都是利用两个预训练模型——GPT-3[7]和稳定扩散模型[51]的互补能力，但不同之处在于，我们利用这些模型生成成对的多模式训练数据。

基于扩散的生成模型 扩散模型[59]方面的最新进展使得最先进的图像合成[10, 18, 19, 53, 55, 60]以及视频[21, 58]、音频[30]、文本[35]和网络参数[45]等其他模式的生成模型成为可能。最近的文本到图像扩散模型[41, 48, 51, 54]已经证明可以从任意文本说明生成逼真的图像。

图像编辑的生成模型 图像编辑模型传统上只针对单一的编辑任务，如风格转换[15, 16] 或图像域之间的翻译[22, 24, 36, 42, 71]。许多编辑方法将图像反转[1-3, 12]或编码[8, 50, 63]到一个潜在空间（如 StyleGAN [25, 26]）中，然后通过操作潜在向量对图像进行编辑。最近的一些模型利用 CLIP [47]嵌入来引导使用文本的图像编辑[5, 9, 14, 28, 31, 41, 44, 70]。我们将 Text2Live [6]与这些方法中的一种进行了比较，这种编辑方法可以优化图像层，使 CLIP 相似性目标最大化。

近期的研究已将预训练的文本到图像扩散模型用于图像编辑[5,17,27,38,48]。虽然有些文本到图像模型本身就具有编辑图像的能力（例如，DALLE-2 可以创建图像变化、修复区域和操作 CLIP 嵌入[48]），但使用这些模型进行有针对性的编辑并非易事，因为在大多数情况下，它们无法保证相似的文本提示会产生相似的图像。Hert 等人最近的研究[17]通过“提示到提示”（Prompt-to-Prompt）方法解决了这一问题，这种方法可以同化相似文本提示生成的图像，从而对生成的图像进行单独编辑。我们在生成训练数据时使用了这种方法。为了编辑非生成图像（即真实图像），SDEdit [38]使用预训练模型对输入图像进行加噪和去噪处理，并添加新的目标提示。我们将 SDEdit 作为基线进行比较。近期的其他研究还包括：根据图片说明和用户绘制的掩膜（mask）进行局部修复[5, 48]；根据一小部分图像生成特定对象或概念的新图像[13, 52]；或通

过反转（和微调）单张图像进行编辑，然后用新的文本描述重新生成图像[27]。与这些方法不同的是，我们的模型只需要一张图像和如何编辑该图像的指令（即不需要任何图像的完整描述），并直接在前向传播中执行编辑，而不需要用户绘制的掩膜、额外的图像或按实例反转或微调。

学习遵循指令 我们的方法不同于现有的基于文本的图像编辑工作[6,13,17,27,38, 52]，它可以根据指令进行编辑，这些指令告诉模型应该执行什么操作，而不是输入/输出图像的文本标签、图片说明或描述。按照编辑指令进行编辑的一个主要好处是，用户可以用自然的书面文字告诉模型该做什么。用户无需提供额外信息，如示例图像或对输入和输出图像之间保持不变的视觉内容的描述。编写的指令具有表现力、精确性和直观性，用户可以轻松分离出需要更改的特定对象或视觉属性。我们的目标是遵循书面图像编辑指令，这一灵感来自于最近的工作，即教大语言模型更好地遵循人类指令完成语言任务[39, 43, 68]。

使用生成模型生成训练数据 深度模型通常需要大量的训练数据。互联网数据集通常是合适的，但可能不以监督所需的形式存在，例如，特定模式的配对数据。随着生成模型的不断改进，人们对将其作为下游任务的廉价而丰富的训练数据来源越来越感兴趣[32, 46, 49, 57, 64, 65]。在本文中，我们使用两种不同的现成生成模型（语言、文本到图像）为我们的编辑模型生成训练数据。

3. 方法

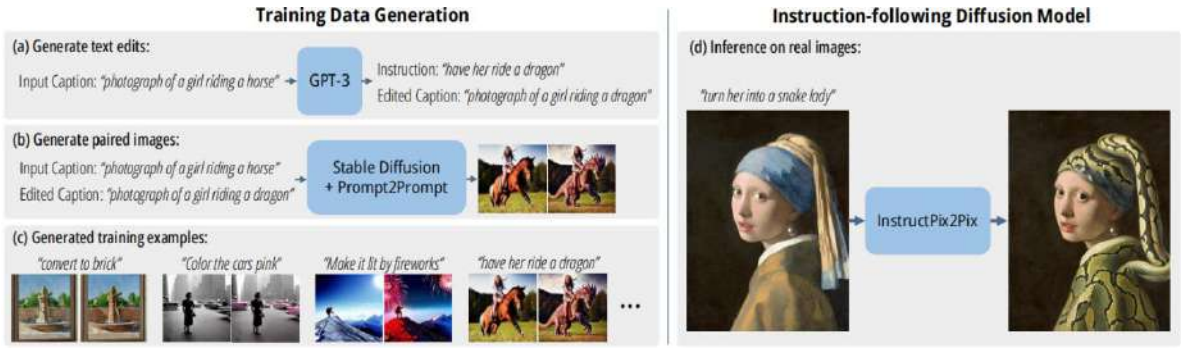

图 2.我们的方法由两部分组成：生成图像编辑数据集和在该数据集上训练扩散模型。(a)我们首先使用经过微调的 GPT-3 生成指令和编辑过的图片说明。(b)然后，我们结合使用稳定扩散模型[51]和 Prompt-to-Prompt [17]，根据成对的图片说明生成成对的图像。我们使用这一程序创建了一个包含超过 450,000 个训练示例的数据集(c)。(d)最后，我们在生成的数据上训练 InstructPix2Pix 扩散模型，以根据指令编辑图像。在推理过程中，我们的模型可以根据人类编写的指令编辑真实图像。

我们将基于指令的图像编辑视为一个监督学习问题：（1）首先，我们生成一个文本编辑指令和编辑前/后图像的配对训练数据集（第 3.1 节，图 2a-c），然后（2）我们在这个生成的数据集上训练一个图像编辑扩散模型（第 3.2 节，图 2d）。尽管我们使用生成的图像和编辑指令进行训练，但我们的模型仍能推广到使用人类编写的任意指令编辑真实图像。我们的方法概览见图 2。

3.1. 生成多模态训练数据集

我们将两个针对不同模态的大规模预训练模型——一个大语言模型[7]和一个文本到图像模型[51]——的能力结合起来，生成了一个多模态训练数据集，其中包含文本编辑指令以及编辑前后的相应图像。在下面两节中，我们将详细介绍这一过程的两个步骤。在第 3.1.1 节中，我们将介绍对 GPT-3 [7]进行微调以生成文本编辑集的过程：给定一个描述图像的提示，生成一个描述要进行修改的文本指令和一个描述修改后图像的提示（图 2a）。然后，在第 3.1.2 节中，我们将介绍使用文本到图像模型[51]将两个文本提示（即编辑前和编辑后）转换成一对相应图像的过程（图 2b）。

3.1.1 生成指令和配对图片说明

	Input LAION caption	Edit instruction	Edited caption
Human-written (700 edits)	*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 generated (>450,000 edits)	*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.我们标注了一个小型文本数据集，对 GPT-3 进行了微调，并使用微调后的模型生成了一个大型文本三联字符集数据集。我们使用 LAION 中的真实图片说明作为标注示例和生成示例的输入说明。高亮文本由 GPT-3 生成。

我们首先完全在文本领域进行操作，利用大型语言模型接收图片说明并生成编辑指令和编辑后的图片说明。例如，如图 2a 所示，如果输入图片说明为 “骑马女孩的照片”，我们的语言模型就能生成 “让她骑龙”这一可信的编辑指令和经过适当修改的输出图片说明“骑龙女孩的照片”。在文本领域的操作使我们能够生成大量不同的编辑集合，同时保持图像变化与文本指令之间的对应关系。

我们的模型是通过在一个相对较小的人工编辑三联数据集上对 GPT-3 进行微调来训练的：（1）输入图片说明，（2）编辑指令，（3）输出图片说明。为了生成微调数据集，我们从 LAION-Aesthetics V2 6.5+ [56]数据集中抽取了 700 个输入图片说明，并手动编写了指令和输出图片说明。我们撰写的指令和输出图片说明示例见表 1a。利用这些数据，我们使用默认训练参数对 GPT-3 Davinci 模型进行了单个轮次的微调。

得益于 GPT-3 丰富的知识和概括能力，我们经过微调的模型能够生成既有创意又合理的指令和图片说明。有关 GPT-3 生成数据的示例，请参见表 1b。我们的数据集是通过使用这个训练有素的模型生成大量编辑和输出图片说明而创建的，其中输入图片说明是 LAION-Aesthetics 的真实图片说明（不包括说明重复或图片 URL 重复的样本）。我们之所以选择 LAION 数据集，是因为它规模庞大、内容多样（包括对专有名词和流行文化的引用）、媒介多样（照片、绘画、数字艺术品）。LAION 的一个潜在缺点是噪音较大，包含大量无意义或无描述性的图片说明——不过，我们发现通过数

据集过滤（第 3.1.2 节）和无分类器引导（第 3.2.1 节）的组合可以减轻数据集噪音。我们最终生成的指令和图片说明语料库包括 454445 个示例。

3.1.2 根据配对字幕生成配对图像

(a) Without Prompt-to-Prompt. (b) With Prompt-to-Prompt.

图 3.使用稳定扩散模型[51]、Prompt-to-Prompt[17]和不使用 Prompt-to-Prompt[17]生成的一对图像。两者对应的图片说明分别为“骑马女孩的照片”和“骑龙女孩的照片”。

接下来，我们使用预训练的文本到图像模型将一对图片说明（指编辑前后的图像）转换成一对图像。将一对图片说明转换成一对相应图像的一个挑战是，文本到图像模型无法保证图像的一致性，即使在条件提示发生非常微小的变化时也是如此。例如，两个非常相似的提示：“猫的图片”和“黑猫的图片”可能会产生完全不同的猫的图像。这不符合我们的目的，因为我们打算将这些配对数据作为训练模型编辑图像的监督数据（而不是生成不同的随机图像）。因此，我们使用了 Prompt-to-Prompt [17]，这是一种最新的方法，旨在鼓励文本到图像扩散模型的多代相似。这是通过在一定数量的去噪步骤中借用交叉注意力权重来实现的。图 3 显示了使用 Prompt-to-Prompt 和未使用 Prompt-to-Prompt 的采样图像对比。

虽然这大大有助于同化生成的图像，但不同的编辑可能需要图像空间中不同的变化量。例如，幅度较大的改变，如改变大规模图像结构的改变（如移动物体、用不同形状的物体替换），可能需要生成的图像对中的相似度较低。幸运的是，Prompt-to-Prompt 有一个可以控制两幅图像相似度的参数：共享注意力权重的去噪步骤ρ的比例。遗憾的是，仅从图片说明和编辑文本中找出ρ的最佳值非常困难。因此，我们为每对图片说明生成 100 对图像样本，每对样本的随机$\rho \sim \mu(0.1, 0.9)$，并使用基于 CLIP 的度量标准对这些样本进行过滤：即 Gal 等人[14]提出的 CLIP 空间中的方向相似性。该指标衡量的是两幅图像（在 CLIP 空间中）之间的变化与两幅图像的图片说明之间变化的一致性。进行这种过滤不仅有助于最大限度地提高图像对的多样性和质量，还能使我们的数据生成在 Prompt-to-Prompt 和稳定扩散模型失败时更加稳健。

3.2. InstructPix2Pix

我们利用生成的训练数据来训练一个条件扩散模型，该模型可根据书面说明编辑图像。我们的模型基于稳定扩散模型（Stable Diffusion），这是一种大规模文本到图像的潜在扩散模型。

扩散模型[59]通过一系列去噪自编码器来学习生成数据样本，这些自编码器会估算数据分布的分数[23]（指向高密度数据的方向）。潜在扩散[51]通过在带有编码器ε和

解码器D的预训练变分自编码器[29]的潜在空间中运行，提高了扩散模型的效率和质量。对于图像x，扩散过程会将噪声添加到编码的潜变量$z=\varepsilon(x)$中，产生一个有噪声的潜变量z_t，噪声水平会随着时间步$t\in T$而增加。我们学习一个网络ϵ_θ，该网络能根据图像条件c_I和文本指令条件c_T预测添加到有噪声潜变量z_t中的噪声。我们要最小化以下潜变量扩散目标：

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

Wang 等人[66]的研究表明，在图像翻译任务中，微调大型图像扩散模型的效果优于从头开始训练模型，尤其是在配对训练数据有限的情况下。因此，我们利用预训练的稳定扩散检查点来初始化模型的权重，充分利用其强大的文本到图像生成能力。为了支持图像调节，我们在第一个卷积层中添加了额外的输入通道，将z_t和$\varepsilon(c_I)$连接起来。扩散模型的所有可用权重都从预训练的检查点初始化，而对新添加的输入通道起作用的权重初始化为零。我们重新使用了原本用于图注的文本调节机制，将文本编辑指令c_T作为输入。其他训练细节见附录 C。

3.2.1 针对两种条件的无分类器引导

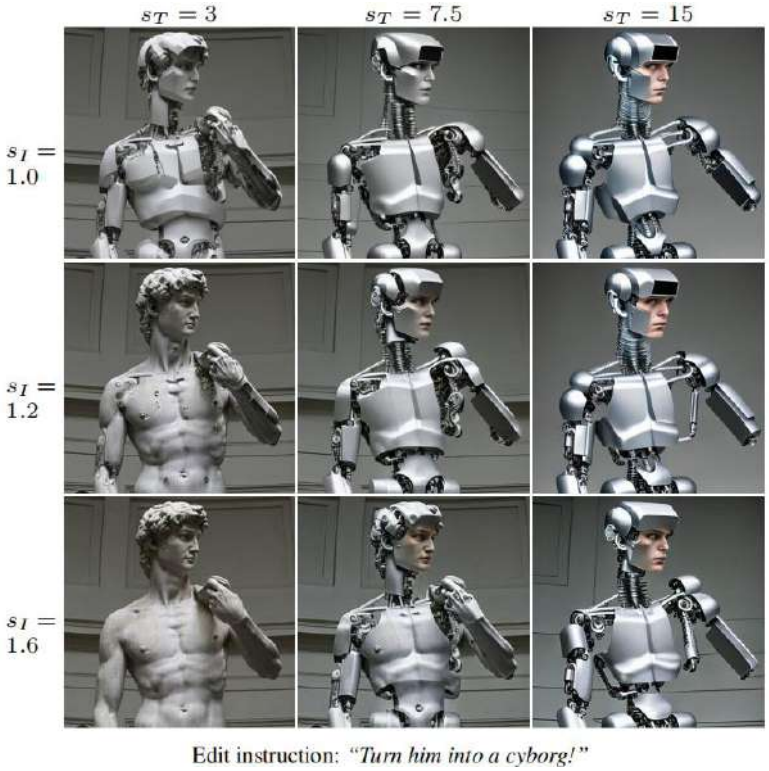

图 4.两个条件输入的无分类器引导权重。s_I值越高，编辑后的图像空间结构与输入图像越相似，而s_T值越高，输出图像的编辑强度越大。

无分类器扩散引导[20]是一种在扩散模型生成的样本质量和多样性之间进行权衡的方法。它通常用于类别条件和文本条件图像生成，以提高生成图像的视觉质量，并使采样图像更好地符合其条件。无分类器引导能有效地将概率质量转移到隐式分类器$p_\theta(c|z_t)$为条件c赋值可能性较高的数据上。无分类器引导的实施涉及对扩散模型进行

有条件和无条件去噪的联合训练，并在推理时将两个分数估计值结合起来。训练无条件去噪时，只需在训练期间的某个频率将条件设置为固定的空值$c=\varnothing$。推理时，在指导尺度$s\geq 1$的情况下，修正后的分数估计值$\tilde{e}_\theta(z_t,c)$会向条件$e_\theta(z_t,c)$的方向外推，并远离无条件$e_\theta(z_t,\varnothing)$。

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

对于我们的任务，得分网络$e_\theta(z_t,c_I,c_T)$有两个条件：输入图像c_I和文本指令c_T。我们发现，在这两个条件下利用无分类器指导是有益的。Liu 等人[37]的研究表明，条件扩散模型可以从多个不同的条件值中得出分数估计值。我们将同样的概念应用到我们的模型中，并使用两个独立的条件输入。在训练过程中，我们对 5%的示例随机设置$c_I=\varnothing_I$，对 5%的示例随机设置$c_T=\varnothing_T$，对 5%的示例同时设置$c_I=\varnothing_I$和$c_T=\varnothing_T$。因此，我们的模型能够对这两种条件输入或其中一种条件输入进行有条件或无条件去噪。我们引入了两个引导尺度：s_I和s_T。增加s_I会使编辑后的图像更接近输入图像，而增加s_T则会使编辑强度更大。我们的修正分数估算如下：

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

图 4 显示了这两个参数对生成样本的影响。有关我们的无分类器指导公式的详细信息，请参阅附录 D。

4. 结果

图 5.《蒙娜丽莎》变身为各种艺术媒介。

图 6.《创造亚当》的新背景和主题（以 768 分辨率生成）

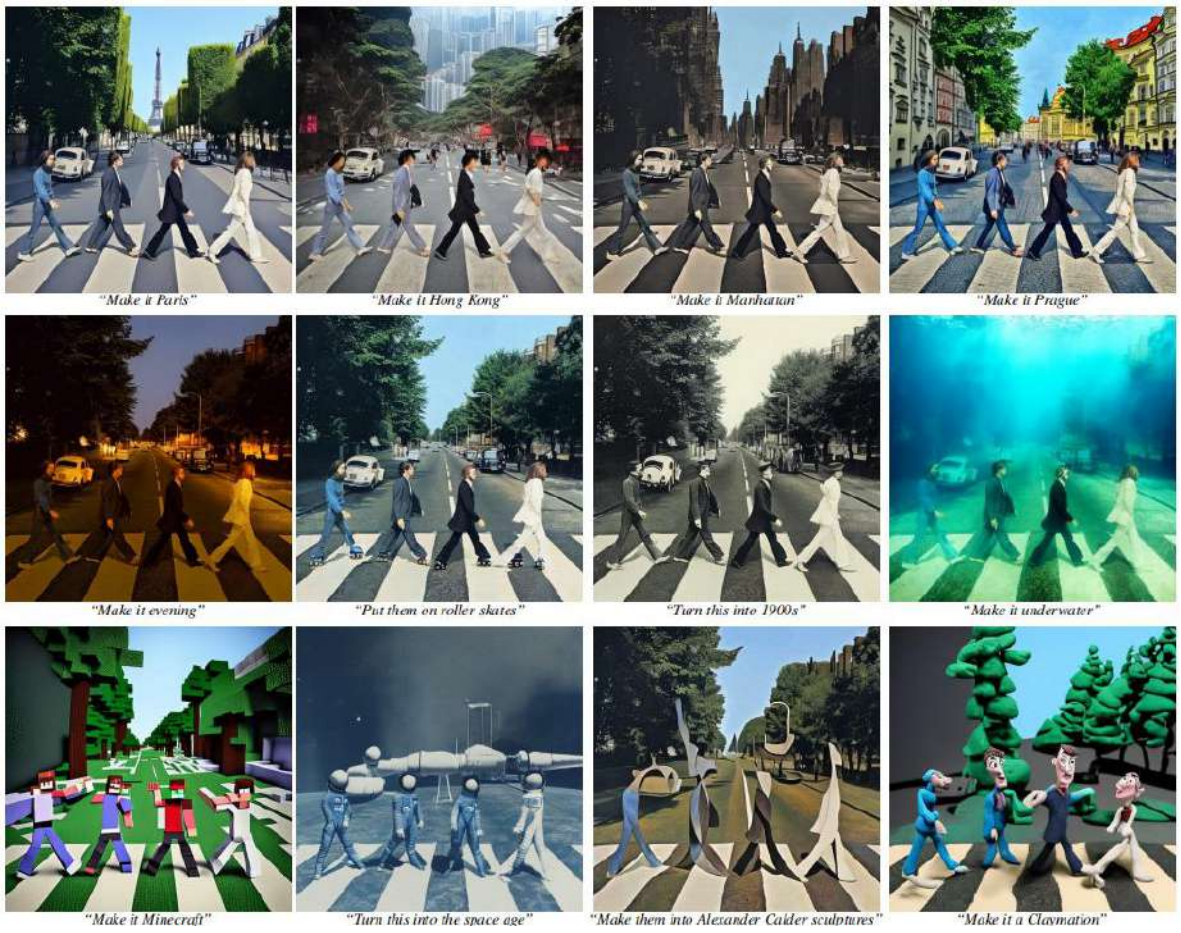

图 7. 披头士乐队标志性的《艾比路》专辑封面以各种方式进行了改造。

我们展示了多种编辑类型和指令下基于指令的图像编辑结果，包括一组真实照片和艺术作品。部分结果见图 1、5、6、7、11、12 和附录 A。我们的模型成功执行了许多具有挑战性的编辑，包括替换对象、改变季节和天气、替换背景、修改材料属性、转换艺术媒介以及其他各种编辑。

我们将我们的方法与最近的工作 SDEdit [38]、Text2Live [6]和 Prompt-to-Prompt[17]进行了定性比较。我们的模型遵循的是如何编辑图像的说明，但之前的作品（包括这些基线方法）期望的是图像（或编辑层）的描述。因此，我们向他们提供的是“编辑后”的文本图像说明，而不是编辑指令。我们还使用两个衡量图像一致性和编辑质量的指标，将我们的方法与 SDEdit 和 Prompt-to-Prompt 进行了定量比较，详见第 4.1 节。最后，我们在第 4.2 节中展示了生成的训练数据的大小和质量如何影响我们模型的性能。

4. 1. 基线比较

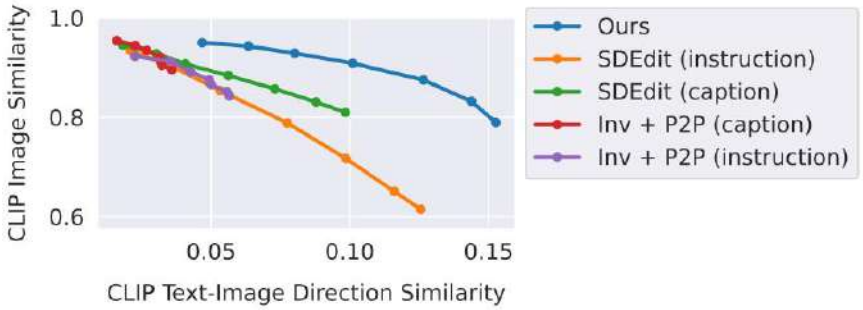

图 8.我们绘制了与输入图像的一致性（Y 轴）和与编辑的一致性（X 轴）之间的权衡图。对于这两个指标，越高越好。我们将文本引导固定为 7.5，并改变：我们方法的$s_I \in [1.0, 2.2]$，SDEdit 的强度（去噪量）为[0.3, 0.9]，Prompt-to-Prompt 的交叉注意周期为[0, 1]。我们使用输出图片说明或编辑指令对 Prompt-to-Prompt 的两种变体进行了实验。

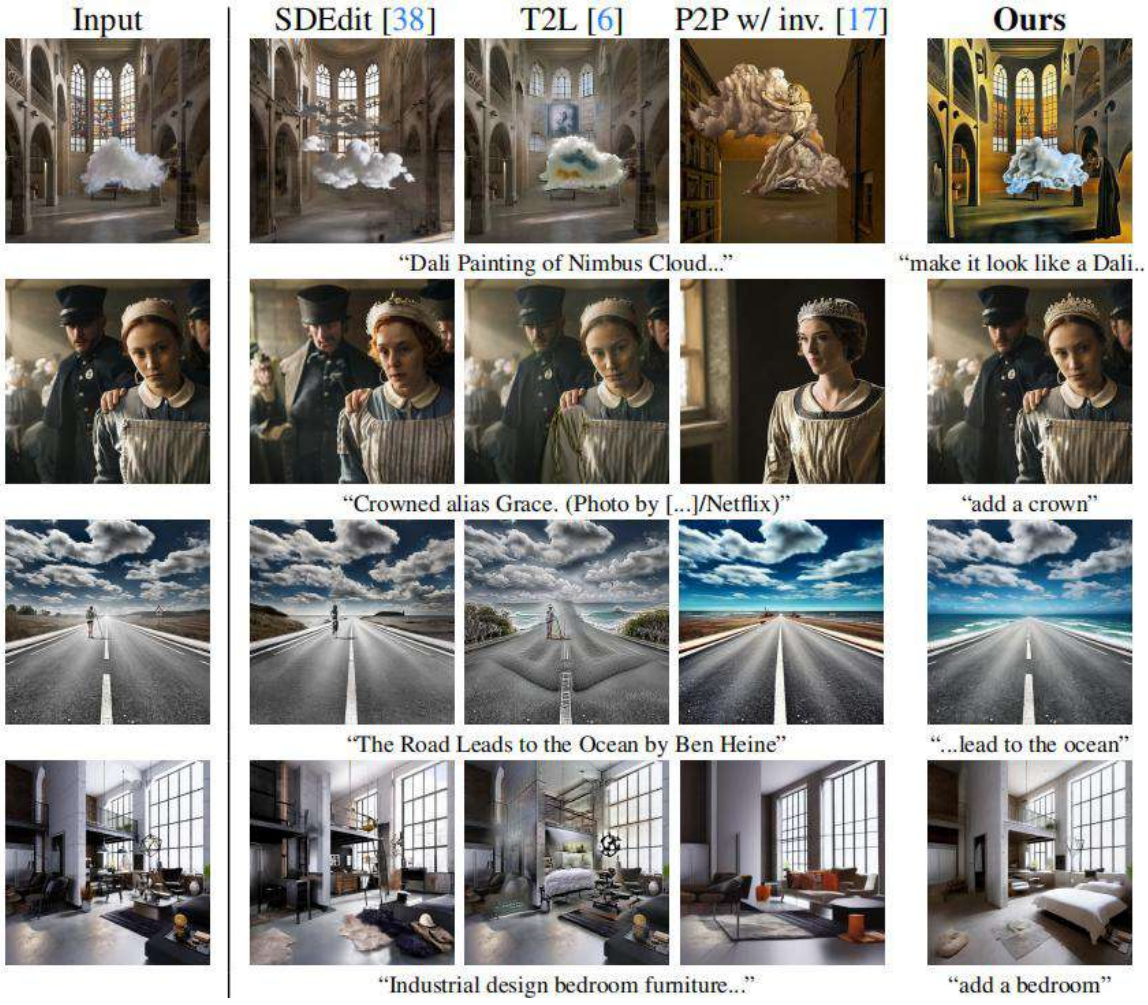

图 9.定性比较。我们与最近的图像编辑方法 SDEdit [38]、Text2Live [6]和 Prompt-to-Prompt [17]进行了比较。这些方法都希望输出图像说明，而我们的方法则不同，它遵循编辑指令。

我们提供了与 SDEdit [38]、Text2Live [6]和 Prompt-to-Prompt [17]的定性比较，以及与 SDEdit 和 Prompt-to-Prompt 的定量比较。SDEdit [38]是一种利用预训练的扩散模型编辑图像的技术，它将部分噪声图像作为输入，然后进行去噪处理，生成新的编辑

图像。Text2Live[6]通过根据文本提示生成颜色+不透明度增强层来编辑图像。

我们在图 9 中对 SDEdit、Text2Live 和 Prompt-to-Prompt 进行了定性比较。其他例子的比较以及这些相关作品的其他配置见附录 B。我们注意到，虽然 SDEdit 在内容大致保持不变、风格有所改变的情况下效果还算不错，但它在保持特性和隔离单个对象方面却很吃力，尤其是在需要进行较大改动时。此外，它需要对所需图像进行完整的输出描述，而不是编辑指令。另一方面，虽然 Text2Live 能够为涉及添加层的编辑提供令人信服的结果，但其表述方式限制了它所能处理的编辑类别。

图 8 显示了与 SDEdit 和 Prompt-to-Prompt 的定量比较。我们绘制了 CLIP 图像嵌入的余弦相似度（编辑后的图像与输入图像的吻合程度）和[14]引入的方向 CLIP 相似度（文字说明的变化与图像的变化的吻合程度）这两个指标之间的权衡图。这是两个相互竞争的指标——提高输出与所需编辑对应的程度会降低输出与输入图像的相似度——我们感兴趣的是哪种方法能实现最佳权衡（最高曲线）。我们发现，与 SDEdit 和 Prompt-to-Prompt 相比，在相同的图像相似度值下，我们的结果实现了更高的方向相似性，这表明它能更好地执行所需的编辑。这些结果是根据 2000 次编辑的平均值测得的，我们还使用不同的 CLIP 模型进一步验证了这些结果（见附图 23）。由于我们在生成训练数据时使用了 Prompt-to-Prompt 模型，因此该模式的表现优于 Prompt-to-Prompt 似乎有些出人意料，我们将在附录 B 中讨论这种改进的可能原因。

4. 2. 消融实验

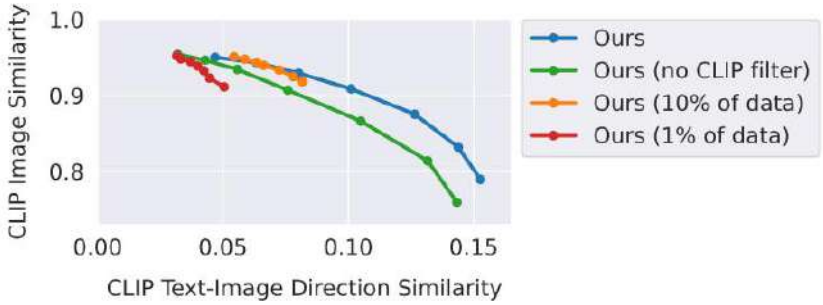

图 10. 我们通过固定s_T和扫频$s_I \in [1.0, 2.2]$，比较了我们模型的消融变体（较小的训练数据集，无 CLIP 过滤）。我们建议的配置表现最佳。

在图 10 中，我们提供了 3.1 节中描述的数据集大小选择和数据集过滤方法的量化消融。减小数据集的大小通常会导致执行更重要的图像编辑的能力下降，而只能执行微妙或风格化的图像调整（因此，图像相似度得分较高，但方向性得分较低）。相比之下，从数据集生成中移除 CLIP 过滤会降低图像与输入图像的整体一致性。

我们还在图 4 中分析了两种无分类器引导尺度的效果。增加s_T会对图像进行更强的编辑（即输出与指令更加一致），而增加s_I则有助于保持输入图像的空间结构（即输出与输入图像更加一致）。我们发现，s_T值在 5-10 范围内和s_I值在 1-1.5 范围内通常能产生最佳结果。在实践中，以及就本文所显示的结果而言，我们发现调整每个示例的引导权重是有益的，这样可以在一致性和编辑强度之间取得最佳平衡。

5. 讨论

图 11. 以不同的指令重复应用我们的模型产生复合编辑。

图 12. 通过改变潜在噪声，我们的模型可以为相同的输入图像和指令生成多种可能的图像编辑。

图 13. 失败案例。从左到右：我们的模型无法执行视角变化，可能会对图像进行不希望看到的过度修改，有时无法隔离指定对象，以及难以重组或交换对象。

我们展示了一种方法，该方法结合了两个大型预训练模型——一个大语言模型和一个文本到图像模型——来生成一个数据集，用于训练扩散模型来遵循书面图像编辑指令。虽然我们的方法能够对图像进行各种引人注目的编辑，包括风格、介质和其他上下文变化，但仍然存在一些局限性。

我们的模型受限于生成数据集的视觉质量，因此也受限于用于生成图像的扩散模型（在本例中为稳定扩散模型[51]）。此外，我们的方法对新编辑的泛化能力以及在视觉变化和文本指示之间建立正确关联的能力也受到了以下因素的限制：用于微调 GPT-3 的人写指示[7]、GPT-3 创建指示和修改图片说明的能力以及 Prompt-to-Prompt [17] 修改生成图像的能力。特别是，我们的模型在计算物体数量和空间推理（例如，“把它移到图像左边”、“交换它们的位置 ”或 “把两个杯子放在桌子上，一个放在椅子上”）方面很吃力，就像在稳定扩散和 Prompt-to-Prompt 中一样。此外，我们还发现，执行多次连续编辑有时会导致伪影累积。失败的例子见图 13。此外，我们的方法所基于的数据和预训练模型中也存在有据可查的偏差。我们的方法所编辑的图像可能会继承这些偏差或引入其他偏差（附图 14）。

除了缓解上述局限性，我们的工作还提出了一些问题，例如：如何按照指令进行空间推理，如何将指令与用户交互等其他调节方式相结合，以及如何评估基于指令的编辑。结合人类反馈（如使用强化学习）是未来工作的另一个重要方向，可以提高我们的模型与人类意图之间的一致性。