Iterative Multi-granular Image Editing using Diffusion Models

K J Joseph, Prateksha Udhayanan, Tripti Shukla, Aishwarya Agarwal, Srikrishna Karanam,
Koustava Goswami, Balaji Vasan Srinivasan
Adobe Research, Bangalore, India
{josephkj,udhayana,trshukla,aishagar,skaranam,koustavag,balsrini}@adobe.com

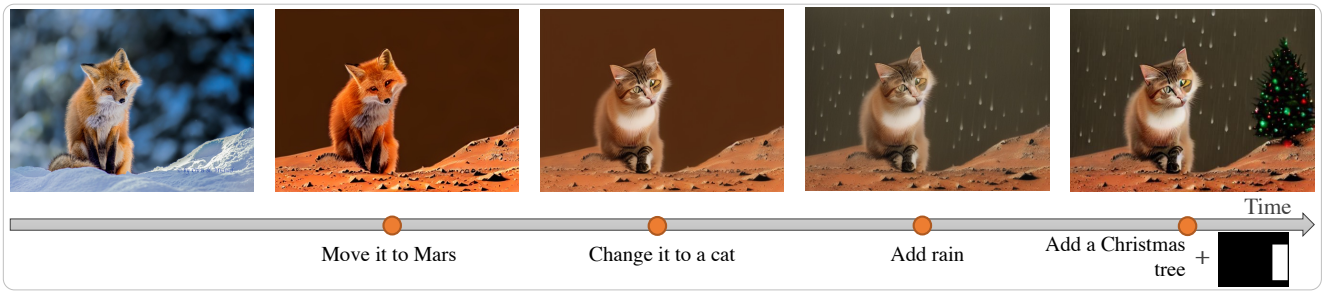

Figure 1. We introduce EMILIE: Iterative Multi granular Image Editor, a diffusion model that can faithfully follow a series of image editing instructions from a user. In this example, we see how the image of the `fox` has been semantically modified according to the provided edit instructions. Optionally, a user can instruct EMILIE with exact location of the desired edit, as illustrated in the last column.

Abstract

Recent advances in text-guided image synthesis has dramatically changed how creative professionals generate artistic and aesthetically pleasing visual assets. To fully support such creative endeavors, the process should possess the ability to: 1) iteratively edit the generations and 2) control the spatial reach of desired changes (global, local or anything in between). We formalize this pragmatic problem setting as `Iterative Multi-granular Editing`*. While there has been substantial progress with diffusion-based models for image synthesis and editing, they are all one shot (i.e., no iterative editing capabilities) and do not naturally yield multi-granular control (i.e., covering the full spectrum of local-to-global edits). To overcome these drawbacks, we propose EMILIE: Iterative Multi-granular Image Editor. EMILIE introduces a novel latent iteration strategy, which re-purposes a pre-trained diffusion model to facilitate iterative editing. This is complemented by a gradient control operation for multi-granular control. We introduce a new benchmark dataset to evaluate our newly proposed setting. We conduct exhaustive quantitatively and qualitatively evaluation against recent state-of-the-art approaches adapted to our task, to being out the mettle of EMILIE. We hope our work would attract attention to this newly identified, pragmatic problem setting.*

1. Introduction

An image is an invaluable form of visual communication. Creating such a visual illustration is a creative process and an expression of the ingenuity of the artist. They often start with a blank canvas and *iteratively* update it with the semantic concepts that they want to convey. These changes might be at different *granularities*, ranging from any small local change to global changes spanning the entire canvas.

Generative technologies for image synthesis have made remarkable strides lately. Diffusion models [8, 21, 26] have had tremendous success in creating realistic images from text prompts. These text-to-image models have the capacity to inspire and augment human creativity. Diffusion models for image editing [1–3, 10, 17, 27, 29] further enhances the collaborative content creation, by allowing users to edit their content using the versatility of diffusion models.

In this work, we identify two major gaps in utilizing diffusion based image editing models as an assistant in a creator's workflow: 1) Current diffusion based image-editing methods are one-shot. They consume an input image, make the suggested edit, and give back the result. This is in stark contrast to the creator's workflow, which is naturally iterative. 2) It is not easy for an artist to specify and constrain the spatial extent of the intended edit. We note that this is a very practical yet under-explored research direction in the literature. Towards this end, we formalize and define this novel

problem setting as *Iterative Multi-granular Image Editing*.

A naive approach to iteratively edit an image would be to use the edited image from the previous step as input to the image editor for the next step. This, however, adds unwanted artifacts to the outputs, and these get accumulated over edit steps as shown in Figs. 2 and 5. To address this, we introduce a new and elegant *latent iteration* framework. Our key insight is that iterating over the latent space instead of the image space, there is a substantial reduction in the amount of noise/artifacts that get added over edit steps. Next, to incorporate multi-granular control, we modulate the denoising process to follow the user-specified location constraints. We interpret the diffusion model as an energy-based model, which allows us to control the spatial extent of the edits by selectively restricting the gradient updates to the regions of interest. These two contributions, put together as part of our overall framework called *EMILIE: Iterative Multi-granular Image Editor*, tackles our newly introduced problem setting. It is critical to note that EMILIE adds these capabilities directly into an already trained diffusion model and does not require any retraining, thereby substantially enhancing its usability.

As we propose and address a novel problem setting, we find that there are no existing benchmark datasets that we can use to evaluate the methodologies. Towards this end, we introduce a new benchmark dataset, called IMIE-Bench, particularly suited to our problem. We extensively evaluate EMILIE both qualitative and quantitatively on our proposed benchmark and show how it performs against various baselines adapted to this new problem. Further, we also evaluate the multi-granular editing capabilities of EMILIE on the publicly available EditBench [29] benchmark. We see that our approach outperforms competing state-of-the-art methods like Blended Latent Diffusion [1] and DiffEdit [3] both qualitatively and quantitatively.

To summarize, the key highlights of our work are:

- We introduce *a novel problem setting* of iterative, multi-granular image editing motivated by practical, real-world use-cases of multimedia content designers.
- To tackle our new problem, we propose EMILIE, a training-free approach that comprises a *new latent iteration method* to reduce artifacts over iterative edit steps and a gradient update mechanism that enables controlling the spatial extent of the edits.
- We propose a *new benchmark dataset*, IMIE-Bench, particularly suited to our new problem setting comprising a carefully curated set of images with a sequence of at least four edit instructions each.
- We conduct *exhaustive experimental evaluation* on IMIE-Bench and EditBench (for multi-granular edits) to bring out the efficacy of our proposed approach, where we clearly outperform the existing state-of-the-art methods.

2. Related Works

Multimodal Image Generation and Editing Methods: Generative Adversarial Networks (GAN) [7] based text to images approaches like StackGAN [33], StackGAN++ [34] and AttnGAN [32] were effective in generating $64 \times 64 \times 3$ images conditioned on their textual description. These methods work well in modelling simple objects like generating flowers and birds, but struggle to generate complex scenes with multiple objects. Recently, diffusion based methods [4, 21] have had phenomenal success in generating realistic images from textual descriptions.

Similar to the success in image synthesis domain, diffusion based models have had significant strides for editing images too. Imagic [10] is able to make complex non-rigid edits to real images by learning to align a text embedding with the input image and the target text. PnP [27], NTI [17], Imagic [10], DiffEdit [3] first use DDIM inversion [4, 26] to invert the image to the input tensor required by the the diffusion model, and then use the text conditioned denoising process to generate the image with the required edit. CycleDiffusion [31] proposes DPM-Encoder to project an image to the latent space of diffusion models thereby enabling them to be used as zero-shot image editors. InstructPix2Pix [2] and Imagen Editor [29] pass the image to be edited directly to the diffusion model by bypassing the DDIM inversion step. This improves the fidelity of the generated edits. But, none of these works addresses the challenges involved with *iterative editing*, which is the main focus of our work. A trivial way to make these efforts iterative would be to pass the edited image recursively. We compare with such iterative variants of PnP [27] and InstructPix2Pix [2] in Sec. 5.

Recent efforts in mask guided image editing methods [1, 3] indeed can support multi-granular editing. We compare with these methods in Sec. 5.

Iterative Image Synthesis: This area is relatively less explored. CAISE [11] and Conversational Editing [16] applies operations like 'rotate image by 90°', 'change the contrast by 60', 'crop the image' and so on. SSCR [6] and Keep drawing it [5] operates on synthetic data from CLEVR [9] dataset, and tries to add objects to a canvas incrementally. Our approach operates on natural images, and can handle real-world edits that an artist might make to an image.

3. Anatomy of Latent Diffusion Models

Diffusion models [8, 25] are a class of probabilistic models that generates a sample from a data distribution, $\boldsymbol{x_0} \sim p(\boldsymbol{x_0})$ by gradually denoising a normally distributed random variable $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, over T iterations. The denoising process creates a series of intermediate samples $\boldsymbol{x}_T, \boldsymbol{x}_{T-1}, \cdots, \boldsymbol{x}_0$; with decreasing amount of noise. The amount of noise to be removed at each step is predicted by a neural network $\epsilon_\theta(\boldsymbol{x}_t, t)$, where t is the denoising step.

These models has been found to be extremely successful in synthesizing realistic images [4, 20, 23] when $\boldsymbol{x}_i \in \mathbb{R}^{W\times H\times 3}$. To reduce the computational requirements for training without degrading quality and flexibility, Rombach *et al*. proposed Latent Diffusion [21], which does the diffusion process in the latent space of a pretrained VQ-VAE [28] encoder. Hence, the input image $\boldsymbol{x}_0$ is first projected into the latent space $\boldsymbol{z}_0 = \mathcal{E}(\boldsymbol{x}_0)$, where $\mathcal{E}(\cdot)$ is the VQ-VAE encoder and $\boldsymbol{z}_0 \in \mathbb{R}^{w\times h\times 4}$, where $w = W/2^4$ and $h = H/2^4$ [21]. The forward diffusion process uses a Markovian noise process q, which gradually adds noise to $\boldsymbol{z}_0$ through $\boldsymbol{z}_T$ with a Gaussian function following a variance schedule β_t:

$$q(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}) = \mathcal{N}(z_t; \sqrt{1-\beta_t}\boldsymbol{z}_{t-1}, \beta_t\mathbf{I}) \tag{1}$$

Ho *et al*. [8] showed that sampling $\boldsymbol{z}_t \sim q(\boldsymbol{z}_t|\boldsymbol{z}_0)$ need not be iterative, but instead can be as follows:

$$q(\boldsymbol{z}_t|\boldsymbol{z}_0) = \mathcal{N}(z_t; \sqrt{\bar{\alpha}_t}\boldsymbol{z}_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{2}$$

$$= \sqrt{\bar{\alpha}_t}\boldsymbol{z}_0 + \epsilon\sqrt{(1-\bar{\alpha}_t)}, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{3}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{r=0}^{t} \alpha_r$. The reverse diffusion process learns $\epsilon_\theta(\boldsymbol{z}_t, t)$ to predict ϵ from Eq. (3). The learning objective is as follows:

$$\mathcal{L}_{LDM} = \mathbb{E}_{t\sim[1,T],\boldsymbol{z}_0=\mathcal{E}(\boldsymbol{x}_0),\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[\|\epsilon - \epsilon_\theta(\boldsymbol{z}_t, t)\|^2] \tag{4}$$

To sample an image from a learned noise prediction function (loosely referred to as diffusion model henceforth) from $\boldsymbol{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, we iteratively apply the following step:

$$\boldsymbol{z}_{t-1} = \boldsymbol{z}_t - \epsilon_\theta(\boldsymbol{z}_t, t) + \mathcal{N}(\mathbf{0}, \sigma_t^2\mathbf{I}) \tag{5}$$

where $\sigma_t = \sqrt{\beta_t}$. We refer the readers to Ho *et al*. [8] for more details on the sampling process. Finally, $\boldsymbol{z}_0$ is projected back to the image space by the VQ-VAE decoder $\mathcal{D}$: $\boldsymbol{x}_0 = \mathcal{D}(\boldsymbol{z}_0)$.

The diffusion model $\epsilon_\theta(\cdot,\cdot)$ is generally implemented with a time-conditional UNet [22] architecture. It contains an encoder, middle layer and a decoder. Each of these modules contains multiple blocks with a residual convolutional layer, self-attention layer and cross-attention layer. The diffusion model can be optionally augmented to consume an extra condition $\boldsymbol{y}$ (like text, caption, canny-maps, segmentation masks, skeletal information *etc*.) as follows: $\epsilon_\theta(\boldsymbol{z_t}, t, \boldsymbol{y})$. The cross-attention layers effectively warps the extra conditioning into the diffusion model.

Latent Diffusion Models (LDM) have been effectively applied in image editing applications too. The major challenge is to make modifications specified via a textual condition $\boldsymbol{y}$ to an existing image $\boldsymbol{x}_0$. Methods like PnP [27], NTI [17], Imagic [10], DiffEdit [3] first invert $\boldsymbol{x}_0$ to $\boldsymbol{z}_{inv}$ using techniques similar to DDIM Inversion [4, 26], so that when they start the diffusion process (in Eq. (5)) from $\boldsymbol{z}_{inv}$, they will be able to get back $\boldsymbol{x}_0$. Alternatively, approaches like InstructPix2Pix [2] and Imagen Editor [29] learn few extra layers to directly pass $\boldsymbol{x}_0$ as input to the diffusion model. This approach retains better characteristics of the input image, as it side-steps the inversion process. We build EMILIE by extending InstructPix2Pix [2] to support iterative and multi-granular control, as it has minimal changes from LDM [21] (the input tensor $\boldsymbol{z}_t$ is augmented with four more channels to consume $\boldsymbol{x}_0$, and the corresponding weights are initialized to zero. No other changes to LDM [21]), and that its trained model is publicly available.

4. Iterative Multi-granular Editing

As previously noted in Section 1, existing techniques are all focused on one-shot generation (no iterative capabilities) and do not provide the ability for users to specify and constrain the spatial extent of the intended edits. Since creative professionals would want to iteratively edit images on the canvas while needing more spatial control on where in the image (global or local) the edits go, the first contribution of this work is the formulation of a new problem setting we call *Iterative Multi-granular Image Editor*. Given an input image $\boldsymbol{I}_0$, a set of edit instructions $E = \{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_k\}$, and an optional set of masks $M = \{\boldsymbol{m}_1, \cdots, \boldsymbol{m}_k\}$ corresponding to each $\boldsymbol{y}_i$, such an image editor $\mathcal{M}(\boldsymbol{I}_i, \boldsymbol{y}_i, \boldsymbol{m}_i)$ should be able to make the semantic modification intended by $\boldsymbol{y}_i$ on $\boldsymbol{I}_i$ iteratively. If the mask $\boldsymbol{m}_i$ is provided by the user, then the model $\mathcal{M}$ should constrain the edits to follow $\boldsymbol{m}_i$. The set of edited images $\mathcal{I}_{edits} = \{\boldsymbol{I}_1, \cdots, \boldsymbol{I}_k\}$ should be visually appealing and semantically consistent to E and M.

Here, we propose one instantiation for $\mathcal{M}(\cdot,\cdot,\cdot)$ that builds on top of a pre-trained diffusion model [2] and does not need any retraining, instead relying only on test-time optimization of the intermediate representation of the model. The practicality of our newly introduced problem setting combined with the versatility of diffusion models allows EMILIE to be a faithful co-pilot for image editing workflows. We explain how EMILIE handles multi-granular control and iterative edits in Sec. 4.1 and Sec. 4.2 respectively. We explain our overall framework in Sec. 4.3.

4.1. Multi-granular Image Editing

The first contribution of our work is to equip diffusion models with the flexibility to constrain where an edit should be applied spatially on an image. We propose to interpret a diffusion model as an energy-based model (EBM) [12, 15], leading to a flexible new method to selectively restrict gradient updates to the region of interest to the user.

We first begin with a brief review of EBMs. An EBM provides a flexible way of modelling data likelihood. The probability density for the latent embedding $\boldsymbol{z} = \mathcal{E}(\boldsymbol{x})$, cor-

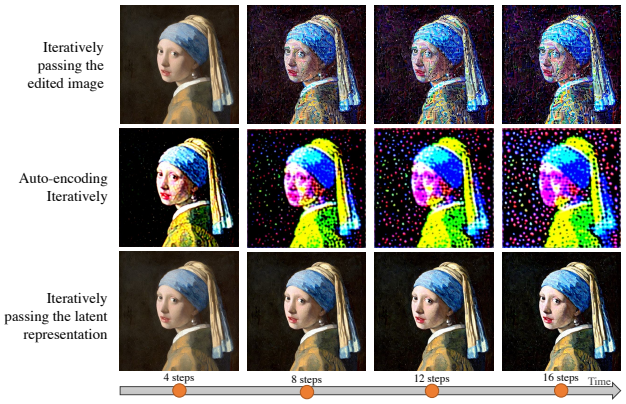

Figure 2. While iteratively passing an image through the diffusion model [2], we see noisy artifacts being added in successive steps (we illustrate $4, 8, 12$ and 16 steps) in Row 1. We see a similar phenomena while only auto-encoding the same image iteratively (here, the diffusion model is not used) in Row 2. Finally, when we iterate in the latent space (where $\boldsymbol{z}_{img}^{e+1}$ is passed iteratively), we see that the successive images are robust to such noisy artifacts.

responding to an image $\boldsymbol{x}$ can be expressed as follows:

$$p_\psi(\boldsymbol{z}) = \frac{\exp\left(-E_\psi(\boldsymbol{z})\right)}{\int_{\boldsymbol{z}} \exp\left(-E_\psi(\boldsymbol{z})\right) d\boldsymbol{z}}, \tag{6}$$

where $E_\psi(\cdot)$ is an energy function which maps $\boldsymbol{z}$ to a single scalar value, called the energy or score. $E_\psi(\cdot)$ is usually instantiated as a neural network with parameters ψ. After learning, sampling from the EBM is indeed challenging, owing to the intractability of the partition function ($\int_{\boldsymbol{z}} \exp\left(-E_\psi(\boldsymbol{z})\right) d\boldsymbol{z}$). A popular MCMC algorithm called Langevin Sampling [18, 30], which makes use of the gradient of $E_\psi(\cdot)$ is used as the surrogate as follows:

$$\mathbf{z}_t = \mathbf{z}_{t-1} - \frac{\lambda}{2} \partial_{\mathbf{z}} E_\psi(\mathbf{z_{t-1}}) + \mathcal{N}(0, \omega_t^2 \mathbf{I}) \tag{7}$$

where λ is step size and ω captures the variance of samples.

Interestingly, the sampling process used by the EBM in Eq. (7) and that used by the latent diffusion model in Eq. (5) are functionally same. Without loss of generality, this allows us to express the *noise predictions* from the diffusion model $\epsilon_\theta(\cdot, \cdot)$ as the *learned gradients* of the energy function $E_\psi(\cdot)$ [15]. Thus we can control the granularity of each edits by selectively zeroing out the gradients (equivalent to masking parts of the noise predictions) that we are not interested in updating. This theoretically grounded approach turns out to be simple to implement. For a user provided mask $\boldsymbol{m} \in \{0, 1\}^{w \times h \times 1}$, which specifies the location to constrain the edits, we update the iterative denoising step in Eq. (5) as follows:

$$\boldsymbol{z}_{t-1} = \boldsymbol{z}_t - \boldsymbol{m} * \epsilon_\theta(\boldsymbol{z}_t, t) + \mathcal{N}(\mathbf{0}, \sigma_t^2 \mathbf{I}) \tag{8}$$

Our experimental analysis in Sec. 5.4, shows that such gradient modulation is effective in localizing the user intent (conveyed via the binary mask), without adding any extra training or compute expense.

4.2. Iterative Image Editing

The next contribution of our work is to iteratively edit images with diffusion models while maintaining the state of the canvas, i.e., the newer edit instruction from a user should be applied on the latest version of the image.

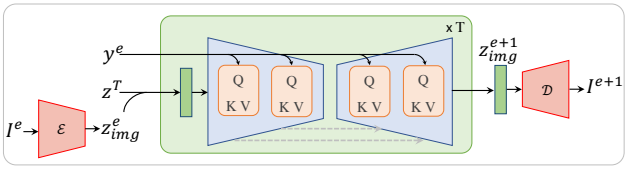

Figure 4. The figure illustrates the architectural components [2] involved in editing an image $\boldsymbol{I}^e$, according to an edit instruction $\boldsymbol{y}^e$ in the e^{th} edit iteration. $\boldsymbol{z}^T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is successively denoised by the diffusion model for T iterations to generate $\boldsymbol{z}_{img}^{e+1}$.

Diffusion models capable of doing image editing [2, 29] can consume the image to be edited. We briefly describe its architectural components in Fig. 4. Consider $\boldsymbol{y}^e$ to be the edit instruction that needs to be applied to an image $\boldsymbol{I}^e$, at an edit step e. $\boldsymbol{I}^e$ is passed though a pretrained VQ-VAE encoder $\mathcal{E}$ to obtain $\boldsymbol{z}_{img}^e$. $\boldsymbol{z}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the initial latent variable. $\boldsymbol{z}_{img}^e$ is stacked with $\boldsymbol{z}_t$ and successively denoised by the pretrained diffusion model (shown in the light green outline) over T iterations.

A naïve approach to iteratively edit the image would be to pass the edited image $\boldsymbol{I}^{e+1}$ through the model along with the next edit instruction $\boldsymbol{y}^{e+1}$. Unfortunately, this accumulates and amplifies the noisy artifacts in the image. The first row of Fig. 2 illustrates this phenomena. In this experiment, we iteratively pass an image through the architecture defined in Fig. 4, and use $\boldsymbol{y}^e = \phi$ (this is to characterize the behaviour independent of the edit instruction at each step).

On a high level, there would potentially be only two sources that might introduce such noisy artifacts: 1) the VQ-VAE based auto-encoder and 2) the denoising steps of the latent diffusion model. In-order to isolate each of their contribution, we experiment using only the VQ-VAE based encoder and decoder from the pipeline described in Fig. 4; *i.e.* $\boldsymbol{I}^{e+1} = \mathcal{D}(\mathcal{E}(\boldsymbol{I}^e))$. We show the corresponding results in the second row of Fig. 2. These result illustrates that the auto-encoding is not perfect, and is indeed accumulating noise over time. This motivates us to propose *latent iteration*, where the latent representation produced by diffusion model corresponding to edit instruction $\boldsymbol{y}^e$ will be passed iteratively along with $\boldsymbol{y}^{e+1}$ in successive edit step. Hence, $\boldsymbol{z}_{img}^{e+1}$ will be used instead of $\mathcal{E}(\boldsymbol{I}^{e+1})$ as input to diffusion model. This simple approach alleviates the exacerbation of the noisy artefacts as seen in the last row of Fig. 2.

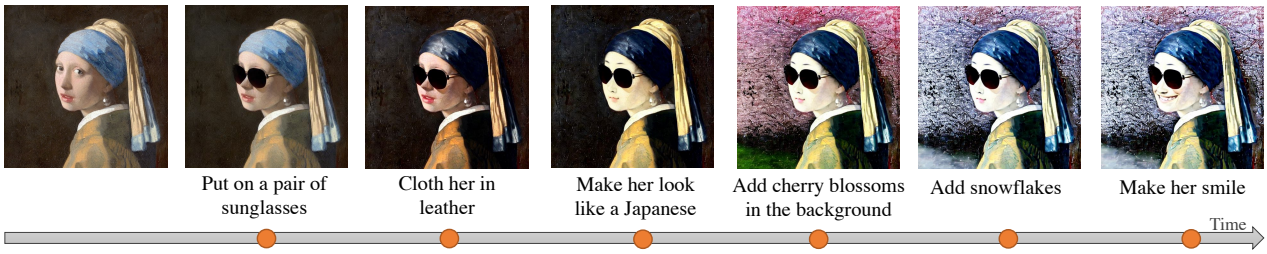

Figure 3. The figure shows how our proposed *latent iteration* framework is able to semantically modify the first image, consistent with the text caption provided by the user at each time step. Please see Sec. 4.2 for more details.

(a) Iteratively passing edited images. (b) Iteratively passing modified latents.

Figure 5. The figure compares the fourth edit step in Fig. 3 between image space and latent space iteration. Green boxes 2, 3 and 5 shows improved preservation of semantic concepts from the previous edit step, while box 4 shows how newer related concepts are added, while keeping the image less noisy (box 1). Kindly zoom in for fine details.

In Fig. 3, we apply latent iteration over multiple edit instructions $\boldsymbol{y}^e$. We note that our method is able to make semantic edits consistent with the corresponding caption in each edit iteration. Further, in Fig. 5, we compare the fourth edit result between iterating over image space and latent space. (We showcase the results of the other steps in the Supplementary owing to space constrains.) It is interesting to note that latent iteration is not only able to reduce the noise accumulation (box 1), but also improve the consistency with previous edits. In box 3, the `sunglass` from the previous step is retained as is (note the frame), while in box 2 and 5, the `leather` head cover is maintained with the newer edits. In box 4, latent iteration was able to add related semantic concept to improve the consistency of the image. We see that this behaviour consistently holds across our exhaustive experimental evaluation in Sec. 5.

4.3. Overall Framework

Algorithm 1 summarizes the key steps involved in adding multi-granular edits to an image iteratively. For the first edit instruction $\boldsymbol{y}_1$, we initialize $\boldsymbol{z}^e_{img}$ to the VQ-VAE encoding of the input image $\boldsymbol{I}_0$ in Line 4. This latent is therein passed to the diffusion model via $\boldsymbol{z}_T$ in Line 7. If the user specifies a mask to control the spatial reach of the edit, it is used while denoising the latent in Step 9. After denoising $\boldsymbol{z}_0$ is indeed decoded using VQ-VAE decoder, and stored to $\mathcal{I}_{edits}$. Importantly, $\boldsymbol{z}_0$ is saved and reused for subsequent edit instructions in Line 11 and Line 6 respectively.

Algorithm 1 Iterative Multi-granular Image Editor

Input: Image to be Edited: $\boldsymbol{I}_0$; Edit Instructions: $E = \{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_k\}$; Optional Masks: $M = \{\boldsymbol{m}_1, \cdots, \boldsymbol{m}_k\}$; Pre-trained Diffusion Model [2]: $\epsilon_\theta(\cdot,\cdot,\cdot)$; VQ-VAE: $\mathcal{E}(\cdot), \mathcal{D}(\cdot)$; Variance Schedule: σ_t; Number of Diffusion Steps: T.
Output: Edited Images: $\mathcal{I}_{edits}$

1: $\boldsymbol{z}_{init} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $\boldsymbol{y}_e \in E$ **do** ▷ *For each edit instruction.*
3: **if** e == 1 **then**
4: $\boldsymbol{z}^e_{img} \leftarrow \mathcal{E}(\boldsymbol{I}_0)$
5: **else**
6: $\boldsymbol{z}^e_{img} \leftarrow$ prev_latent ▷ *Latent Iteration.*
7: $\boldsymbol{z}_T \leftarrow$ concat$(\boldsymbol{z}_{init}, \boldsymbol{z}^e_{img})$
8: **for** $t \in \{T, \cdots, 0\}$ **do** ▷ *For each denoising step.*
9: $\boldsymbol{z}_{t-1} = \boldsymbol{z}_t - \boldsymbol{m}_e * \epsilon_\theta(\boldsymbol{z}_t, t, \boldsymbol{y}_e) + \mathcal{N}(\mathbf{0}, \sigma_t^2 \mathbf{I})$ ▷ *Eq.* (8)
10: $\mathcal{I}_{edits}$.append$(\mathcal{D}(\boldsymbol{z}_0))$
11: prev_latent = $\boldsymbol{z}_0$
12: **return** $\mathcal{I}_{edits}$

5. Experiments and Results

5.1. IMIE-Bench Benchmark

To complement our newly introduced problem setting, we introduce IMIE-Bench (Iterative Multi-granular Image Editing Benchmark) to evaluate the efficacy of the methodologies. Inspired by TEdBench [10], we manually curate 40 images from LAION [24] dataset. These include images of people, landscapes, paintings, and monuments. Next, we collected four semantically consistent edit instructions that would modify these images. This would help to quantify the quality of the iterative edits by the model. We also add masks to some of these edit instructions to simulate local editing. We hope that IMIE-Bench would serve as a standardized evaluation setting for this task.

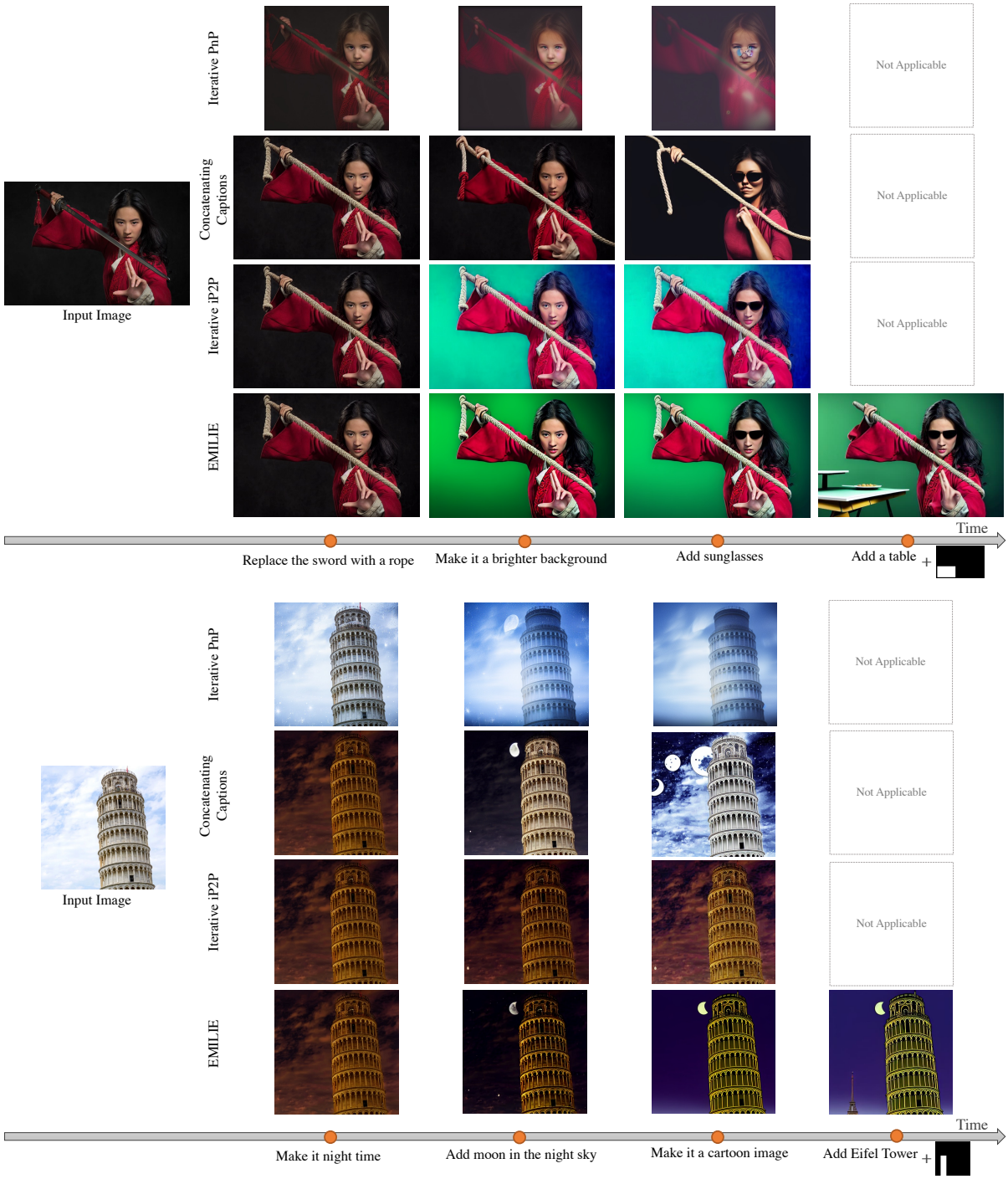

Figure 6. Here we compare EMILIE with three baselines on images from IMIE-Bench. 'Iterative PnP' and 'Iterative iP2P' refers to Plug and Play [27] and Instruct Pix2Pix [2], where the latest edited image is recursively passed on in the next edit step. 'Concatenating Captions' refers to the setting where all the edit instructions that have been received so far, is concatenated and send through Instruct Pix2Pix [2] to edit the original image. We see that EMILIE is able to consistently reduce the noisy artefact accumulation and retain semantics of earlier steps better. Please see more results in the Supplementary material.

5.2. Experimental Protocol

We evaluate the 'iterative' and 'multi-granular' aspects of our approach both qualitatively and quantitatively. We use examples from IMIE-Bench and compare against three baseline approaches. Two of these include iteratively passing the edited image through two image editing methods: InstructPix2Pix [2] and Plug and Play [27]. These methods

are representative of two kinds of approaches for diffusion-based image editors. Plug and Play uses DDIM inversion to project the image to the latent space of the diffusion model, while InstructPix2Pix learns a few new layers to directly pass the input image into the model. The third baseline is to concatenate all the instructions received until a time step, and pass it through InstructPix2Pix along with the original image. We show these results in the following sections.

Further, to better understand the multi-granular local editing aspect of EMILIE, we compare with two recent state-of-the-art mask-based image editing methods: DiffEdit [3] and Blended Latent Diffusion [1] on the EditBench [29] dataset. We could not compare with Imagen Editor [29] because their code, models or results are not publicly available. We use two complementary metrics for quantitative evaluation: first, we compute the CLIP [19] similarity score of the edited image with the input edit instruction and the second is a text-text similarity score, where BLIP [13] is used to first caption the generated image, and it is then compared with the edit instruction.

5.3. Implementation Details

We build EMILIE by extending the InstructPix2Pix diffusion model [2] to support iterative multi-granular editing capability. We keep all the hyper-parameters involved to be same as their implementation. We use a single NVIDIA A-100 GPU to do our inference. Each edit instruction takes 6.67 seconds on average to complete. While doing latent iteration, we find that the magnitude of values in $\boldsymbol{z}_{img}^{e+1}$ is significantly lower than $\mathcal{E}(\boldsymbol{I}^{e+1})$. To normalize this, we multiply $\boldsymbol{z}_{img}^{e+1}$ with a factor $f = avg(\mathcal{E}(\boldsymbol{I}^{e+1}))/avg(\boldsymbol{z}_{img}^{e+1})$. We use ancestral sampling with Euler method to sample from the diffusion model.

5.4. Results on IMIE-Bench

We showcase our qualitative results in Fig. 6. We compare with iterative versions of InstructPix2Pix [2] and Plug and Play (PnP) [27] and a concatenated set of edit instructions, on images from IMIE-Bench.

While analyzing the results, we note that EMILIE is able to retain the concepts from the previous edits, without deteriorating the quality in successive iterations. Iterative PnP struggles the most. This can be attributed to the DDIM inversion which projects the image to the latent space of the diffusion model. The concatenated caption (that contains all the edit instruction that we have seen so far) contains multiple concepts. The diffusion model tries its best to generate these multi-concept images, but struggles to maintain consistency with the previously edited version of the image. Iterative InstructPix2Pix performs the closest to EMILIE, but accumulates noise as edits progresses.

EMILIE supports multi-granular edits too, as is shown in the last column. The baseline methods cannot operate in this setting. We compare EMILIE with multi-granular approaches in Sec. 5.5. These results illustrate that iterating in latent space indeed helps attenuate noise, and gives the model the optimal plasticity to add new concepts yet retain the stability to not forget the previous set of edited concepts.

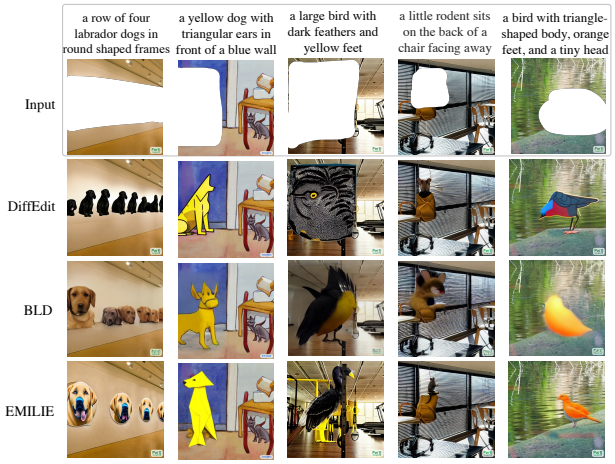

Figure 7. We illustrate the ability of EMILIE to do local editing on images from EditBench [29] benchmark. We compare against recent state-of-the-art methods, DiffEdit (ICLR '23) [3] and Blended Latent Diffusion (SIGGRAPH '23) [1]. Our simple gradient modification approach is able to consistently improve the quality of generation when compared to these approaches.

5.5. Results on EditBench

Fig. 7 shows the results of using EMILIE for doing localized edits in images. The top row includes the input image with mask and the corresponding edit text from the user. The subsequent two rows include results from two recent state-of-the-art approaches: DiffEdit [3] and Blended Latent Diffusion [1], followed by EMILIE. The results show that the methods are successful in limiting the extent of edits to the area constrained by the user. EMILIE is able to make semantically richer modifications to the masked region. Our training-free guidance during the denoising steps is able to constrain the edits to local regions, while being more semantically consistent.

Finally, we run a quantitative evaluation of how well the edited image is able to capture the semantics of the edit instruction successfully by measuring the CLIP and BLIP scores. The results are provided in Tab. 1. We comfortably outperform the baselines here too.

5.6. User Study

We conduct a user study with images from IMIE-Bench and EditBench [29] datasets to test the preference of users across the generated images in Tab. 3 and Tab. 2 respectively. The former evaluates iterative edits, while the latter analyzes the preference for multi-granular edits. From the

Table 1. We quantitatively evaluate the performance of EMILIE for multi-granular editing here. When compared to recent state-of-the-art approaches, EMILIE is able to score better performance in both CLIP and BLIP score metrics.

	DiffEdit [3]		BLD [1]		EMILIE
Average CLIP Score	0.272	(-0.039)	0.280	(-0.031)	**0.311**
Average BLIP Score	0.582	(-0.038)	0.596	(-0.024)	**0.620**

Table 2. User Study for Multi-granular Edits.

Method	Split
DiffEdit [3]	4.87%
BLD [1]	13.33%
EMILIE	81.79%

Table 3. User Study for Iterative Edits.

Method	Split
Concat	18.18 %
iP2P	19.32 %
EMILIE	62.50 %

30 users, most of them preferred the generations from EMILIE over the other baselines.

6. Further Discussions and Analysis

6.1. Object Insertion with EMILIE

Our proposed approach gives control to the user in specifying the location of edits. At times, users might have an image of an object that they would like to insert into a base image. It would be great if the model can automatically propose a plausible location and then insert the object there. We find a straight forward strategy to combine EMILIE with GracoNet [35] (a recent method for proposing object placement in images) to achieve this. Given a base image and an inset image, GracoNet proposes masks for potential location for the inset image. Next, we pass the inset image thought BLIP 2 [13] model to generate a caption. Finally, the mask, the caption and base image is passed to EMILIE to render the final image.

Fig. 9 showcases some examples from OPA dataset [14]. We can see that the object insertion by EMILIE is more composed and realistic than that by GracoNet. This is because, instead of just placing the object at a location, EMILIE indeed synthesises a new object at the location specified via the mask proposed by GracoNet.

6.2. Ablating the Gradient Control

In order to understand the contribution of our proposed gradient control strategy explained in Sec. 4.1, we turn it off while doing local editing. Results in Fig. 8 shows that without modulating the latent representations selectively (following Eq. (8)), the edit instructions gets applied globally.

6.3. Limitations

While experimenting with EMILIE, we could understand that it cannot handle negative edit instructions. Let us say we add a pair of sun-glasses as the first edit, and try removing it in the second edit step, the model fails to give back the original image. Disentangling the feature representations for each edits would potentially help to alleviate this issue. We will explore this in future work. We show more failure cases in the Supplementary materials.

Figure 8. We re-purpose the local editing capability of EMILIE to insert objects into specific locations of a source image. For this, we combine EMILIE with GracoNet [35] which predicts the location of the object to be inserted. We can see from the results that the insertions done by EMILIE is more composed to the background.

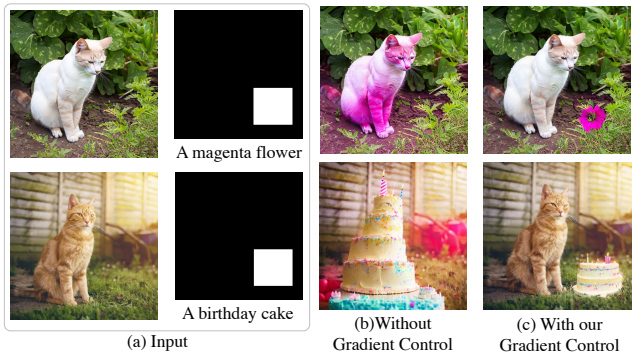

Figure 9. Here we experiment with turning off the gradient control strategy explained in Sec. 4.1 for local edits. Without modulating the gradients, the characteristics of the edit instruction gets globally applied on the image, whereas EMILIE is able to easily localise the edits effectively.

7. Conclusion

We introduce a novel problem setting of *Iterative Multi-granular Image Editing* where a creative professional can provide a series of edit instructions to be made to a real image. Optionally, they can specify the spatial locality on which the edit needs to be applied. Our proposed approach EMILIE is training free and utilizes a pre-trained diffusion model with two key contributions: latent iteration (Sec. 4.2) for supporting iterative editing and gradient modulation (Sec. 4.1) for supporting multi-granular editing. Finally, we introduce a new benchmark dataset IMIE-Bench, and bring out the mettle of our approach by comparing EMILIE against other state-of-the-art approaches adapted to our novel task. We hope that this newly identified research area would be actively investigated by the community.

References

[1] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM Transactions on Graphics (TOG)*, 2023. 1, 2, 7, 8

[2] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022. 1, 2, 3, 4, 5, 6, 7

[3] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 2, 3, 7, 8

[4] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 2, 3

[5] Alaaeldin El-Nouby, Shikhar Sharma, Hannes Schulz, Devon Hjelm, Layla El Asri, Samira Ebrahimi Kahou, Yoshua Bengio, and Graham W Taylor. Keep drawing it: Iterative language-based image generation and editing. *arXiv preprint arXiv:1811.09845*, 2, 2018. 2

[6] Tsu-Jui Fu, Xin Eric Wang, Scott Grafton, Miguel Eckstein, and William Yang Wang. Sscr: Iterative language-based image editing via self-supervised counterfactual reasoning. *arXiv preprint arXiv:2009.09566*, 2020. 2

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1, 2, 3

[9] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. 2

[10] 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. 1, 2, 3, 5

[11] Hyounghun Kim, Doo Soon Kim, Seunghyun Yoon, Franck Dernoncourt, Trung Bui, and Mohit Bansal. Caise: Conversational agent for image search and editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10903–10911, 2022. 2

[12] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and Fujie Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006. 3

[13] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 7, 8

[14] Liu Liu, Zhenchen Liu, Bo Zhang, Jiangtong Li, Li Niu, Qingyang Liu, and Liqing Zhang. Opa: object placement assessment dataset. *arXiv preprint arXiv:2107.01889*, 2021. 8

[15] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, pages 423–439. Springer, 2022. 3, 4

[16] Ramesh Manuvinakurike, Trung Bui, Walter Chang, and Kallirroi Georgila. Conversational image editing: Incremental intent identification in a new dialogue task. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 284–295, 2018. 2

[17] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. *arXiv preprint arXiv:2211.09794*, 2022. 1, 2, 3

[18] Radford M Neal et al. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011. 4

[19] 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. 7

[20] 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. 3

[21] 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

[22] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 3

[23] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 3

[24] 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

[25] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 2

[26] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1, 2, 3

[27] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. *arXiv preprint arXiv:2211.12572*, 2022. 1, 2, 3, 6, 7

[28] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 3

[29] Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J Fleet, Radu Soricut, et al. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. *arXiv preprint arXiv:2212.06909*, 2022. 1, 2, 3, 4, 7

[30] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. pages 681–688, 2011. 4

[31] Chen Henry Wu and Fernando De la Torre. Unifying diffusion models' latent space, with applications to cyclediffusion and guidance. *arXiv preprint arXiv:2210.05559*, 2022. 2

[32] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018. 2

[33] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017. 2

[34] Zizhao Zhang, Yuanpu Xie, and Lin Yang. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6199–6208, 2018. 2

[35] Siyuan Zhou, Liu Liu, Li Niu, and Liqing Zhang. Learning object placement via dual-path graph completion. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, pages 373–389. Springer, 2022. 8