Context Diffusion: In-Context Aware Image Generation

Ivona Najdenkoska[1,2*] Animesh Sinha[2] Abhimanyu Dubey[2] Dhruv Mahajan[2]
Vignesh Ramanathan[2] Filip Radenovic[2]
[1]University of Amsterdam [2]GenAI, Meta

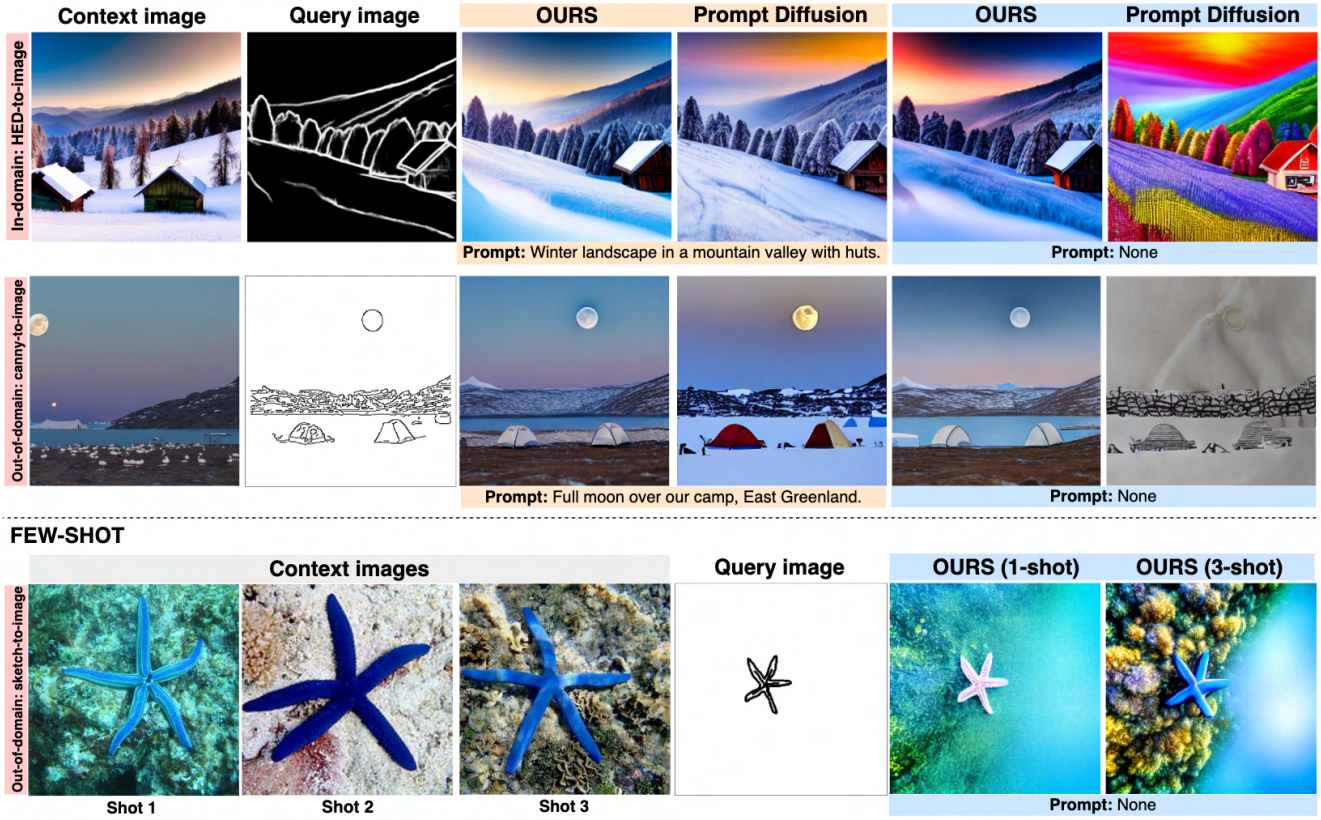

Figure 1. **Illustrating in-context aware image generation with Context Diffusion.** **Top row**: HED-to-image as an in-domain task; **Middle row**: canny-to-image as an out-of-domain task. Our model enables learning from context with and without prompts. The counterpart model, Prompt Diffusion [43] does not leverage the context if the prompt is not provided, hinting at its over-reliance on textual guidance; **Bottom row**: Few-shot setting for sketch-to-image task. More context examples help in learning stronger visual cues, even without prompts.

Abstract

We propose Context Diffusion, a diffusion-based framework that enables image generation models to learn from visual examples presented in context. Recent work tackles such in-context learning for image generation, where a query image is provided alongside context examples and text prompts. However, the quality and fidelity of the generated images deteriorate when the prompt is not present, demonstrating that these models are unable to truly learn from the visual context. To address this, we propose a novel framework that separates the encoding of the visual context and preserving the structure of the query images. This results in the ability to learn from the visual context and text prompts, but also from either one of them. Furthermore, we enable our model to handle few-shot settings, to effectively address diverse in-context learning scenarios. Our experiments and user study demonstrate that Context Diffusion excels in both in-domain and out-of-domain tasks, resulting in an overall enhancement in image quality and fidelity compared to counterpart models.

*Work done during an internship at GenAI, Meta.

1. Introduction

Generative models are witnessing major advances, both in natural language [6, 9, 27, 40, 45, 49] and media generation [5, 18, 28, 33, 36]. Large language models in particular have shown impressive in-context learning capabilities [6, 46]. This is the ability of a model to learn from a few samples on the fly, without any gradient-based updates, and extend it to new tasks. However, for generative models in computer vision, learning from context examples is still under-explored.

Prompt Diffusion [43] is perhaps the closest line of work that explicitly supports a single source-target image pair as a context example for image generation. It builds on the popular ControlNet [48] model which introduced the idea of visually controllable diffusion models for image generation. Specifically, Prompt Diffusion attempts to learn the visual mapping from the source image to the target context image and applies it to a new query image, by also using a prompt for text-based guidance. However, we empirically observed that this model struggles to leverage the image pairs when the text prompt is absent. This results in low fidelity to the visual context examples, particularly when the examples are from a different domain than what is seen during training. For instance, if the source-target pairs show specific styles, they cannot be learned during inference just from the context examples. This is seen in the first row of Figure 1 where Prompt Diffusion is unable to learn the "snowy" style from the context unless prompted through text. Additionally, it does not trivially support multiple source-target images as context examples, which limits the visual information that can be provided to the model.

We address these challenges with our proposed *Context Diffusion* model that can (i) effectively learn from visual context examples as well as follow text prompts and (ii) support a variable number of context examples since visual characteristics can be defined with more than a single example. Unlike Prompt Diffusion, our model does not require paired context examples, but just one or more "target" context images serving as examples of the desired output and a single query image providing visual structure. The reason for using target examples as context is that the source images are derived from the target itself and do not provide any additional information for the task. Typically, the query image provides guidance for the output structure through edges, depth, segmentation maps, etc. On the other hand, the context images provide hints for finer details like style, texture, colors, and object appearances desired in the output image.

It is important to note the difficulty in controlling both aspects of the output image solely through the control mechanism used in a ControlNet-like model. The "control" part of the model is very effective in capturing high-level structure. However, finer details are better captured through the conditioning mechanism. A similar observation is made in previous works such as textual inversion [12] and retrieval-augmented image generation [4, 7], where object appearance is preserved by encoding it through conditioning. Inspired by this, we inject information from the context images into the network in a similar fashion as text conditioning. In particular, we sum the visual embeddings from the context images and place them alongside the text embeddings in the cross-attention layers of the diffusion model. This allows stronger reliance on the visual input from the context and also supports multiple context images. The structure from the query image is preserved by passing it as a control signal to the network in a similar manner as ControlNet [48].

We follow a similar training strategy as Prompt Diffusion [43], by learning from six different tasks using generated images and their maps. At inference time, we use a query image to define the target structure and one or more context images to provide finer visual signals, alongside an optional text prompt. Our experiments study the generation ability of Context Diffusion for in-domain tasks, such as using HED, segmentation, and depth maps to generate real images and vice versa. We show the flexibility of our model to preserve structure from the query image and transfer other visual signals from the context even when the text prompt is missing. Moreover, to properly study in-context learning abilities, we experiment with unseen, out-of-domain tasks, such as handling sketches as query images, image editing, and more. This demonstrates the generalization abilities of Context Diffusion, unlike previous works. Furthermore, for such tasks, using multiple images as context helps improve the fidelity of the generated images to the context.

Contributions. (i) We propose Context Diffusion, an in-context aware image generation framework. It enables pre-trained diffusion models to use visual context examples to control the appearance of the output image, alongside a query image that defines structure and an optional text prompt. (ii) We enable the use of multiple context images as "few-shot" examples for image generation. To the best of our knowledge, this is the first work to explore such a "few-shot" setup for in-context aware image generation. (iii) We conduct extensive offline and online (human) evaluations that show that our framework can handle several in-domain and out-of-domain tasks and demonstrates improved performance over the counterpart model.

2. Related Work

Diffusion-based Image Generation. Recent advancements in diffusion models, first introduced in [38] have exhibited huge success in text-to-image generation tasks [10, 17, 31, 32, 36]. Enhancements have been achieved through various training [10, 33, 36] and sampling [24, 39, 44] techniques. For instance, DALLE-2 [32] proposed an architec-

ture encompassing several stages, by encoding text with CLIP [30] language encoder and decoding images from the encoded text embeddings, followed by Imagen [36] which showed that up-scaling the text encoder largely improves the text fidelity. Furthermore, the Latent Diffusion Model (LDM) [33] investigated the diffusion process by applying it to a low-resolution latent space and even further improved the training efficiency. However, all these models only take a text prompt as input, which restricts the flexibility of the generation process as it requires extensive prompt engineering to obtain the desired image outputs.

Controllable Image Generation. Adding more control to the image generation process, besides the text prompts, helps in overall customization and task-specific image generation. Recent text-conditioned models focus on adjusting models by task-specific fine-tuning [8, 12, 35], injecting conditioning maps, like segmentation maps, sketches or keypoints [2, 3, 11, 23, 29, 48], or exploring editing abilities [5, 13, 15, 25, 26]. For instance, SpaText [2] is using segmentation maps where each region of interest is annotated with text, to better control the layout of the generated image. Models like GLIGEN [22] inject grounding information, such as bounding boxes or edge maps, into new trainable layers via a gated cross-attention mechanism. ControlNet [48], as a recent state-of-the art in controllable image generation presents a general framework for adding spatial conditioning controls. UniControl [29] extends ControlNet by unifying various image map conditions into a single image generation framework. Other works, such as Re-Imagen [7] and RDM [4], employ retrieval for choosing images given a text prompt, for controlling the generation process.

Our approach differs from these models in several aspects. We support in-context learning from visual examples as an addition to the textual prompts and query images. This allows learning new tasks using the visual context only, which yields a more flexible framework. Additionally, we use only a few of the image maps considered by ControlNet and UniControl for training, namely HED, segmentation, and depth maps, and demonstrate the generalization ability to the other visual controls *i.e*. query images.

In-Context Learning in Image Generation. In-context learning is vastly explored both in language-only [6, 20, 45, 46] and visual-language models [1, 19, 21, 41], as an emergent ability enabling to learn new tasks without additional gradient-based updates. However, the ability to learn from context examples is lagging behind in image generation. Prompt Diffusion [43] presents such a framework, by extending the control abilities of ControlNet [48] and training for in-context image generation. They consider a vision-language prompt encompassing a source-target image pair and a text prompt, which is used to jointly train the model on six different tasks. However, Prompt Diffusion only shows good performance when both the context images and prompt are present. In case the text prompt is not present, the model exhibits deteriorating performance, suggesting its inability to learn efficiently from the visual examples, as shown in Figures 1, 3, 4, 5, 6, and 7.

Different from them, we aim to develop a model able to generate images of good quality even when only one of the conditions (visual context or text prompt) is present, both for in-domain and out-of-domain tasks. Another work tackling image generation with visual examples is Prompt-Free Diffusion [47]. It focuses only on having an image as a context, ie, a visual condition, while completely removing the ability to process textual prompts. This is the major difference compared to our Context Diffusion, since we aim to support both scenarios: having the context images and/or text prompts. Additionally, none of these related works consider settings with multiple examples in context, namely, few-shot scenarios. We propose a framework that can handle a variable number of context images, helpful for enriching the visual context representation.

3. Methodology

3.1. Preliminaries

Diffusion models are a class of generative models that convert Gaussian noise into samples from a learned data distribution via an iterative denoising process. In the case of diffusion models for text-to-image generation, starting from noise z_t, the model produces less noisy samples $z_{t-1}, \dots, z_0$, conditioned on caption representation $\mathbf{c}$ at every time step t.

To learn such a model f_θ parameterized by θ, for each step t, the diffusion training objective $\mathcal{L}$ solves the denoising problem on noisy representations z_t, as follows:

$$\min_\theta \mathcal{L} = \mathrm{E}_{z,\epsilon\sim\mathcal{N}(0,1),t}\big[\, \| \epsilon - f_\theta(z_t, t, \mathbf{c}) \|_2^2 \big], \qquad (1)$$

With large-scale training, the model f_θ is trained to denoise z_t based on text information as the main source of control.

To enable more control over the generation process, we follow the ControlNet setup [48], for encoding the structure of the desired output via a query image as visual control. Note that, according to in-context learning parlance, we use *query image* interchangeably with *visual control*. In this paper, we extend the $\mathbf{c}$ representation in Eq. (1), by adding image examples as additional guidance besides the text prompt. Namely, we inject visual embeddings obtained by a pre-trained vision encoder f_{img} with fixed parameters, in a similar fashion as the prompt embeddings.

3.2. Context Diffusion Architecture

The model f_θ is essentially a UNet architecture [34], with an encoder, a middle block, and a skip-connected decoder.

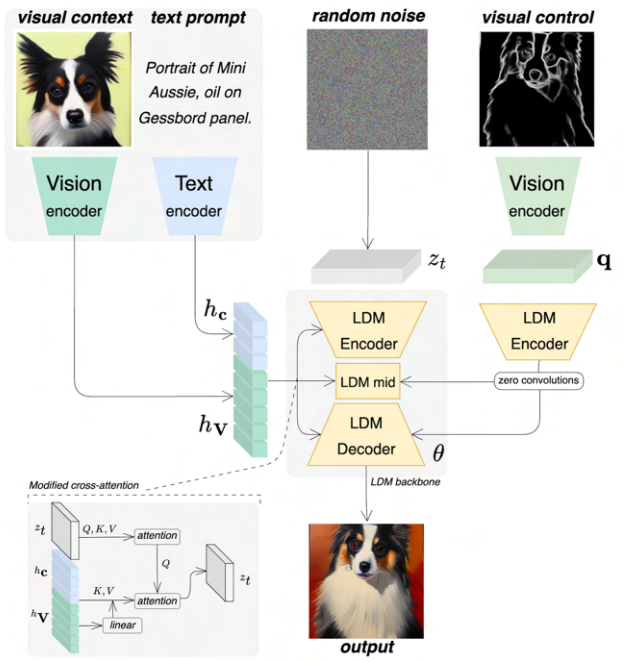

Figure 2. **Architecture of Context Diffusion.** It consists of several modules: vision and text encoders for encoding the text prompt and visual context respectively, an LDM backbone for handling the image generation process, and an additional LDM encoder for processing the query image as a visual control. Note that here we show one visual context example, however, the model is trained using a variable number of such examples.

These modules denoted as LDM encoder, mid, and decoder in Figure 2, are built out of standard ResNet [14] and Transformer blocks [42] which contain several cross-attention and self-attention mechanisms. The core of conditional-diffusion models is encoding the conditional information [33], based on which z_t is generated at a given time step t. We differentiate two types of such conditional information: the visual context $\mathbf{V}$ encompassing images and the text prompt $\mathbf{c}$, to define our conditioning information: $\mathbf{y} = (\mathbf{c}, \mathbf{V})$, where $\mathbf{V} = [\mathbf{v}_1, \dots \mathbf{v}_k]$ and k are the number of images. Additionally, we consider a visual control image, *i.e.*, the query image that serves to define the structure of the output denoted as $\mathbf{q}$.

Prompt encoding. To perform the encoding of the textual prompt $\mathbf{c}$ we use a pre-trained language encoder f_{text} with fixed parameters to obtain the embeddings. Particularly, we obtain $\mathbf{h^c} = \{h_0^\mathbf{c}, \dots, h_{N^\mathbf{c}}^\mathbf{c}\} = f_{\text{text}}(\mathbf{c})$, where $N^\mathbf{c}$ is the number of text tokens, $h_i^\mathbf{c} \in \mathbb{R}^{d^\mathbf{c}}$ and $d^\mathbf{c}$ is the dimensionality of the textual token embeddings.

Visual context encoding. We hypothesize that the visual context $\mathbf{V}$ should be at the same level of conditioning as the textual one. Therefore, we follow a similar strategy for encoding the visual context, by using a pre-trained, fixed image encoder f_{img}. Given a visual context $\mathbf{V}$ consisting of k-images, we encode each image $\mathbf{v}_i$ as $\mathbf{h}^{\mathbf{v}_i} = \{h_0^{\mathbf{v}_i}, \dots, h_{N^\mathbf{v}}^{\mathbf{v}_i}\} = f_{\text{img}}(\mathbf{v}_i)$, where $N^\mathbf{v}$ is the number of tokens per image, $h^{\mathbf{v}_i} \in \mathbb{R}^{d^\mathbf{v}}$ and $d^\mathbf{v}$ is the dimensionality of the visual token embeddings. The final representation of the visual context is obtained by simply summing the corresponding visual tokens of all k-images, where $k \in \{1, 2, 3\}$, yielding $\mathbf{h^V} = \sum_{i=1}^{k} \mathbf{h}^{\mathbf{v}_i}$. Additionally, we add a linear projection layer to map the visual embedding dimension $d^\mathbf{v}$ to the language dimension $d^\mathbf{c}$.

Modified cross-attention. Given the standard cross-attention block in LDMs, defined with queries Q, keys K, and values V, the noisy representation z_t is used as a query, whereas the text encoding $\mathbf{h^c}$ is used as a representation of the keys and values, as follows:

$$z_t = z_t + \text{CrossAtt}(Q = z_t, K = V = \mathbf{h^c}). \quad (2)$$

Our framework is slightly different from this definition since we also consider visual information in the conditioning. Therefore, after obtaining both visual and textual embeddings we simply concatenate them to obtain $[\mathbf{h^c}, \mathbf{h^V}]$, illustrated in the bottom left corner of Figure 2. We hypothesize that the visual and textual conditioning should be at the same level, thus the input to the cross-attention block in (2) changes as follows:

$$z_t = z_t + \text{CrossAtt}(Q = z_t, K = V = [\mathbf{h^c}, \mathbf{h^V}]). \quad (3)$$

Visual control encoding. To enable the ingestion of the query image as visual control, we follow ControlNet setup [48]. First, the image is encoded using a few convolutional layers. Then, a copy of the LDM encoder is used to process the encoded query image $\mathbf{q}$. This trainable LDM encoder copy is connected to the original LDM backbone using zero convolution layers, as shown in Figure 2.

3.3. Multi-task Training Procedure

We use a pre-trained image generation model to adapt it with visual context injection. We use the original denoising objective defined in (1), with $\mathbf{q}$ being the query image and the modified conditioning information $\mathbf{y}$:

$$\min_\theta \mathcal{L} = \mathrm{E}_{z, \epsilon \backsim \mathcal{N}(0,1), t}\big[\parallel \epsilon - f_\theta(z_t, t, \mathbf{y}, \mathbf{q}) \parallel_2^2 \big]. \quad (4)$$

To train with this objective, we use a collection of tasks for joint end-to-end training, similar to [43]. Different from them, we use a visual context sequence consisting of a k-images and a text prompt, together with a query image. Specifically, k is randomly chosen at batch construction. The goal of such training is to leverage any visual characteristics from the context images and to apply them along with the text prompt to the query image.

Prompt dropout. We aim to achieve learning from the context images by avoiding over-reliance on the text prompts. Starting from a pre-trained text-to-image model means the ability to generate images given a text prompt is already strong. Therefore, to enforce the model to pick up cues from an additional conditioning signal *i.e.* the visual context, we apply random replacement of 50% of the text prompts with empty strings, similar to [48], which empirically showed to be an important step.

4. Experiments

4.1. Experimental Setup

Datasets. To train our model, we use a dataset that consists of 310k synthetic images and caption pairs, similar to Prompt Diffusion [43]. Following their training setup, we extract three image maps: HED, segmentation, and depth maps from the training images. During training, for map-to-image tasks, the image maps serve as queries, and real images are used for visual context, while for image-to-map tasks, the real images serve as queries, and image maps are used for visual context. Note that the prompts and visual context are related and describe the same conditioning signal.

At inference time, we use the test partition of the dataset to test the ability to learn from context. To demonstrate the generalization abilities of Context Diffusion to out-of-domain tasks, we extract other image maps, such as normal maps, canny edges, and scribble maps. Also, we consider editing tasks by using real images as queries. To further test the generalization abilities, we use hand-drawn sketches from the Sketchy dataset [37], where the sketch is the query image, and the real images are the context. This dataset does not provide captions, therefore we construct text prompts using a template: "A professional, detailed, high-quality image of *object name*", following [48].

Implementation details. The backbone of our model follows a vanilla ControlNet architecture, initialized in the same way as [48]. We train such a model using the data setup explained above. In particular, only the encoder of the LDM backbone is kept frozen and its copy which processes the query image is trained. For the encoding of the context images and prompts, we use frozen CLIP ViT-L/14 [30] encoders. We take the last-layer hidden states as representations of both the context images and prompts. The model is trained with a fixed learning rate of $1e$-4 for 50K iterations, using 256×256 images. We use a global batch size of 512 for all runs. At inference time, we apply DDIM [39] as a default sampler, with 50 steps and a guidance weight of 3. Regarding the computational resources, the model is trained using 8 NVIDIA A100 GPUs.

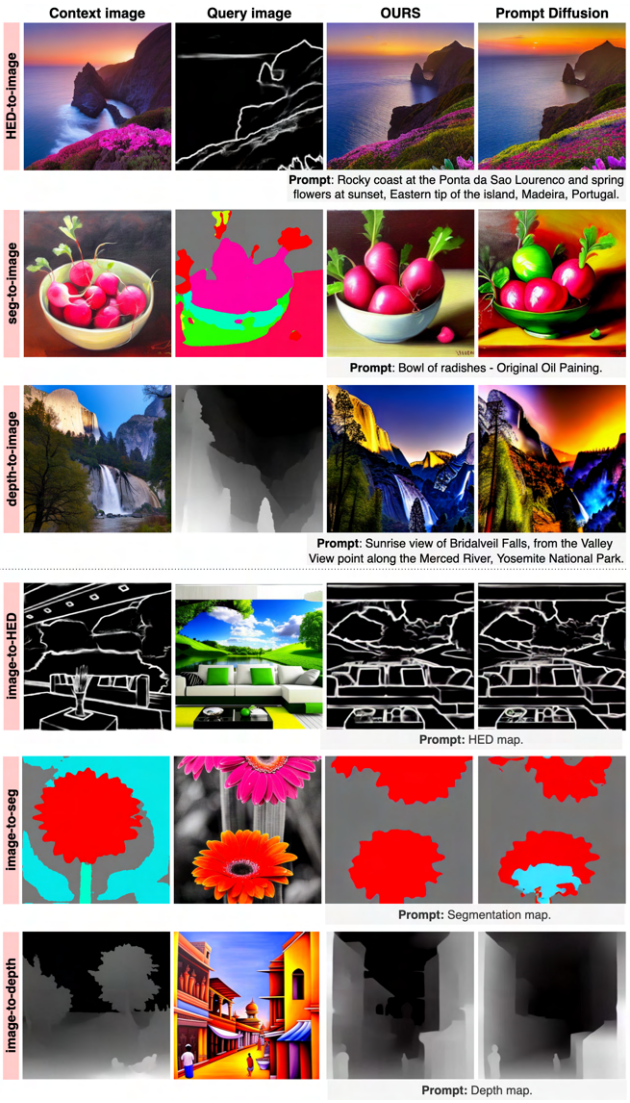

Figure 3. **In-domain tasks comparison to Prompt Diffusion [43]:** Examples of {HED, segmentation, depth}-to-image as forward tasks and image-to-{HED, segmentation, depth} as reverse tasks, with visual context and prompt (C+P) given as conditioning.

User study setup. To better quantify the performance of our model, we perform an online evaluation in the form of a user study to compare our model to Prompt Diffusion [43]. A total of 10 in-house annotators participated in the study, annotating 240 test samples. We present two generated images from the models, randomly annotated as A and B, alongside the given visual context, query image, and prompt. Then, each annotator chooses either a preferred image or both as equally preferred. We consider various in-domain and out-of-domain tasks for evaluation, across three distinct scenarios: using both visual context and prompts, using only visual context and only prompt. Considering these scenarios examines to what extent the models can learn

from the conditional information in a balanced manner and whether they suffer when one input modality is not present.

Automated metrics. In addition to the user study, we also use offline automated metrics to further evaluate the performance of our model. In particular, we report FID scores [16] for map-to-image tasks and RMSE scores for the image-to-map tasks. We use 5000 test images per setting for each task, generated by our model and Prompt Diffusion.

4.2. Results & Discussion

In this section, we compare our model against the most similar approach in the literature, *i.e*. Prompt Diffusion [43]. Prompt Diffusion expects a source-target pair of context images as an input, while in contrast our approach only requires context *i.e*. target images. More analysis regarding this is provided in the supplementary materials. It is important to notice that in all comparisons we follow the source-target format of the input for Prompt Diffusion output image generation, but we omit the visualization of the source image from the Figures, to have consistent visualizations for both methods. Additionally, both approaches operate with a query image and textual prompt as additional inputs.

We compare the methods across two important generalization axes: (i) *in-domain* for seen and *out-of-domain* for unseen tasks at training; (ii) visual context and prompt *(C+P)*, context-only *(C)*, and prompt-only *(P)* variations of conditioning. Finally, we present the results of our model on *few-shot* setup when several context examples are given as input. Prompt Diffusion does not support the few-shot setup.

4.2.1 Data Domain

In-domain comparison. We study the performance of models on the same data domain as the training data, but on the test data that is set aside. This encompasses three "forward" tasks, *i.e*., the query image is either HED, segmentation, or a depth map while the expected output image is a real image, given the visual context and prompt in an adequate form. Similarly, we evaluate three "reverse" tasks, where the query and output roles are reversed. For the purpose of this discussion, we focus on the conditioning setup where both visual context and prompts *(C+P)* are given as input. Figure 3 presents representative examples for each of the tasks: the first three rows depict forward tasks, while the last three rows depict reverse tasks. It can be observed that our model is able to generate images with better fidelity to the context images and prompts, by managing to match the specific colors and styles from the context. On the other hand, Prompt Diffusion outputs are more saturated and fail to leverage the visual characteristics from the context (green radish instead of red in the second row). We include more examples in the supplementary materials. These observations are further supported by user study presented in Table 1 (In-domain (C+P) column), as well as in offline metrics comparison presented in Table 3 (C+P columns), where we obtain satisfactory performance improvement (36.3% *vs*. 28.5% win-rate) over Prompt Diffusion.

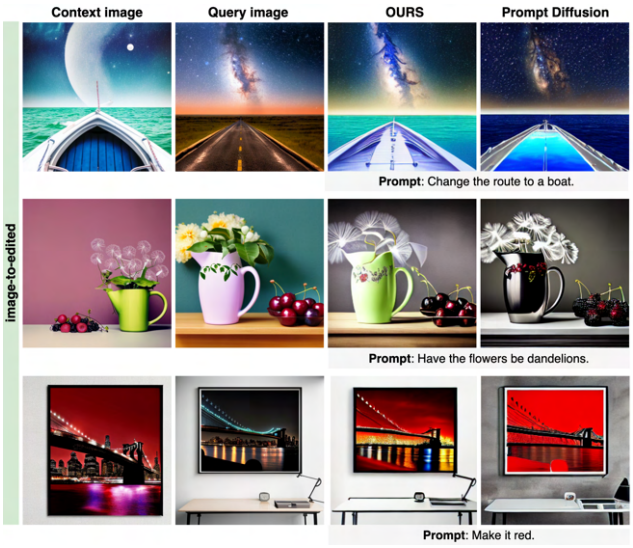

Figure 4. **Out-of-domain comparison to Prompt Diffusion [43]:** Image edit, with visual context and prompt (C+P) as conditioning.

Out-of-domain generalization. The most advantageous aspect of having a model that is an in-context learner is its capacity to generalize to new tasks by observing the context examples given as input at inference. Again, for the purpose of the discussion in this section, we focus on the conditioning setup where both visual context and prompt *(C+P)* are given as input. To test these generalization abilities, we consider tasks outside of training domains: image editing with representative examples in Figure 4; {sketch, normal map, scribbles, canny edge}-to-image with representative examples in Figure 5. In both figures, we observe noticeable improvements over Prompt Diffusion [43]. It is apparent that the visual characteristics of the context images are also transferred in the output images. Furthermore, we select editing and sketch-to-image as representative out-of-domain tasks to perform a user study. We report the results in Table 1 (Out-of-domain (C+P) column), where we observe great improvements in win rate (52.3% *vs*. 26.9%), significantly higher than for in-domain setup, showing the advantage of in-context aware image generation.

4.2.2 Conditioning at inference

Using only visual context. To better understand the effect of visual context examples on the model's performance, we

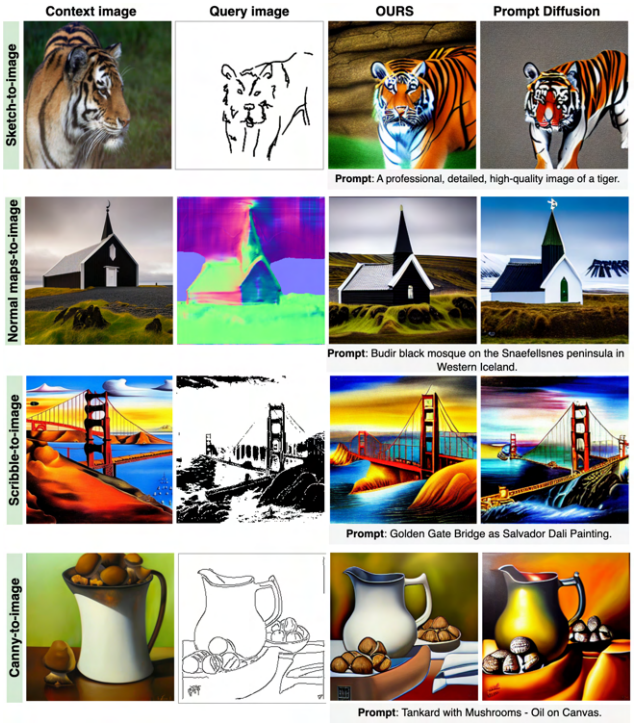

Figure 5. **Out-of-domain comparison to Prompt Diffusion [43]:** {sketch, normal map, scribble, canny edge}-to-image tasks. Visual context and prompt (C+P) are given as conditioning information.

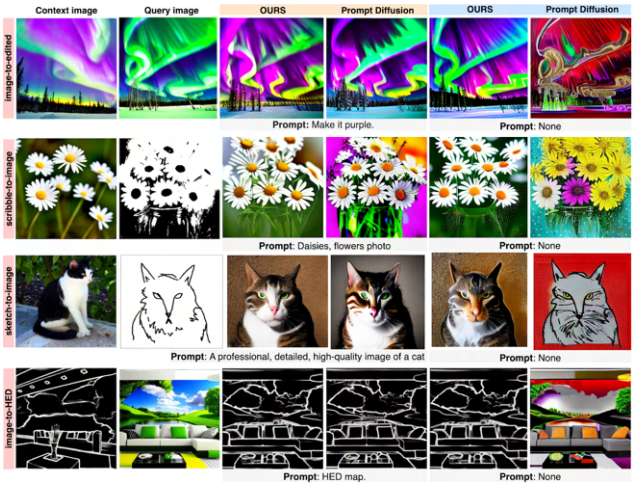

Figure 6. **Conditioning comparison with Prompt Diffusion [43]:** Using visual context and prompt (C+P) and visual context-only (C) as conditioning, on both in-domain (image-to-HED) and out-of-domain (editing, scribble-to-image, sketch-to-image) tasks.

analyze the outputs when the text prompt is not provided (empty string), *i.e.*, only visual context is used as conditioning. This experiment gives strong insights into the model's ability to perform in-context learning. We show representative examples of this setup in Figure 6. It can be observed that Prompt Diffusion [43] is unable to learn from the visual

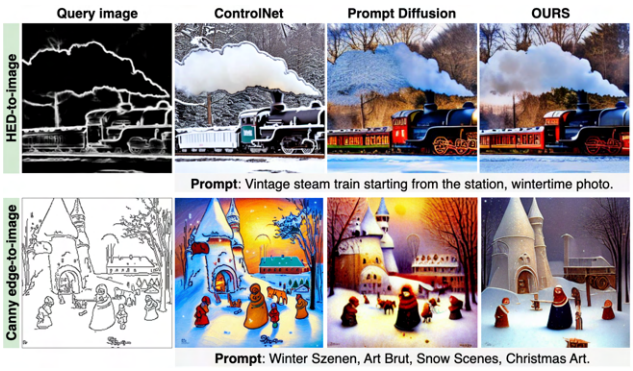

Figure 7. **Zero-shot comparison to ControlNet [48] and Prompt Diffusion [48]:** Using prompt-only (P) as conditioning.

	In-domain				Out-of-domain			
	C+P	C	P	*avg*	C+P	C	P	*avg*
PD [43]	28.5	4.5	**30.4**	21.1	26.9	22.8	25.9	25.2
Ours	**36.3**	**80.2**	29.6	**48.6**	**52.3**	**63.7**	**49.8**	**55.2**

Table 1. **User study comparison to Prompt Diffusion (PD) [43]:** In-domain and out-of-domain tasks, considering different conditioning settings: context image and prompt (C+P), visual context-image-only (C), prompt-only (P). We report the win rate as a percentage of winning votes for each model.

	Out-of-domain		
	C+P	C	*avg*
Ours (1-shot)	21.5	28.3	24.9
Ours (3-shot)	**60.0**	**50.2**	**55.1**

Table 2. **User study comparison for 1-shot *vs*. 3-shot setups:** Out-of-domain tasks, considering different conditioning settings: visual context and prompt (C+P), visual context-only (C). Note that the "prompt-only" (P) setting corresponds to a zero-shot scenario and is not applicable here. We report the win rate as a percentage of winning votes for each model.

examples, indicating that it relies solely on the text caption as conditional information. We include this setting in the user study and we report the results in Table 1 ((C) columns). Overall, we observe a *significant* performance gap between our model and Prompt Diffusion, both for in-domain (80.2% *vs*. 4.5% win-rate) and out-of-domain (63.7% *vs*. 22.8% win-rate) tasks. This result is additionally supported by the offline metrics in Table 3 ((C) columns) for in-domain tasks, further strengthening the observations that our model is able to truly leverage the visual context.

Using only text prompts. Apart from being able to handle scenarios only with visual context, we aim to also support scenarios using only text prompts. To enable this setting, we simply mask out the visual context by using black images.

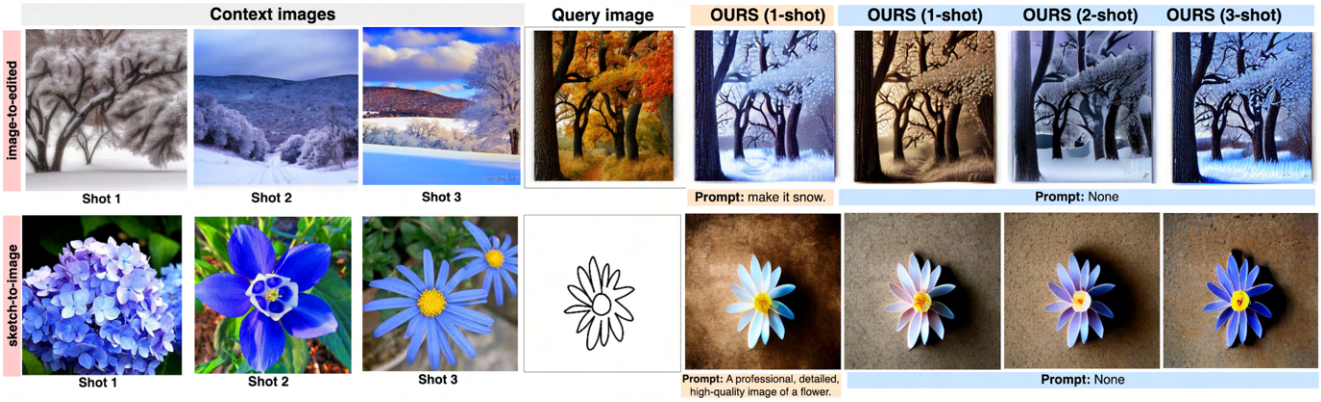

Figure 8. **Few-shot examples:** Comparison between out-of-domain tasks (editing and sketch) using one context example with a text prompt, and one, two, and three shots of context examples with no text prompt. Our model is able to leverage multiple visual examples to handle scenarios when the text prompt is not present.

	FID (map-to-img) ↓									RMSE (img-to-map) ↓								
	HED-to-img			seg-to-img			depth-to-img			img-to-HED			img-to-seg			img-to-depth		
	C+P	C	P	C+P	C	P	C+P	C	P	C+P	C	P	C+P	C	P	C+P	C	P
PD [43]	12.8	22.5	15.1	16.7	25.1	**17.2**	15.9	27.0	18.1	0.15	0.33	**0.15**	0.32	0.41	0.32	**0.14**	0.34	0.14
Ours	**12.3**	**17.7**	**14.8**	**13.4**	**19.0**	18.5	**12.9**	**18.5**	**17.5**	**0.11**	**0.11**	0.16	**0.29**	**0.28**	**0.30**	**0.14**	**0.13**	**0.13**

Table 3. **Offline comparison to Prompt Diffusion (PD) [43] using automated metrics: FID and RMSE:** In-domain tasks across three different conditioning settings: visual context and prompt (C+P), visual context-only (C), prompt-only (P). Lower scores are better.

This essentially yields zero-shot settings, boiling down to how ControlNet [48] is used at inference time. However, unlike ControlNet which requires a separate model trained for each task, our Context Diffusion generalizes across a series of tasks. In Figure 7 we show representative examples of this setting, comparing our model to ControlNet and Prompt Diffusion. It can be seen that our model is able to generate more realistic images compared to ControlNet and also performs on par with Prompt Diffusion. Similar as before, we also include this setting in the user study, reported in Table 1 ((P) columns). We observe that our performance is slightly worse to Prompt Diffusion on in-domain (29.6% *vs*. 30.4%) and much better on out-of-domain (49.8% *vs*. 25.9%) tasks. This supports our observations that the Prompt Diffusion approach relies too much on the textual prompt, as well as suffers in out-of-domain data regimes. Further, we compare the automated metrics in Table 3 ((P) columns), again observing comparable performance on in-domain tasks.

4.2.3 Few-shot visual context examples

The Context Diffusion architecture is flexible enough to accommodate multiple context examples, enabling few-shot scenarios. Using one context example proved to be enough for in-domain tasks, as seen in Figure 3. Therefore in the few-shot experiments, we focus on the out-of-domain tasks, such as editing and sketch-to-image. In particular, we augment the visual context sequence with additional images, depicting similar objects or scenes. Moreover, we look at scenarios where the textual information is not present since in that case, the model has to rely on the visual context only. As can be seen from Figure 8, adding more context examples helps to strengthen the conditional visual representation, especially when the prompt is not present. We also quantify the performance by conducting a user study for the few-shot settings, presented in Table 2. We are comparing our model when using one context example *vs*. using three examples. Overall we observe improved performance (55.1% *vs*. 24.0% average win rate) when using three context images which aligns with the qualitative observations. Note that in the current experiments, we use 1 up to 3 shots as a representative few-shot setting, however, our model can accommodate more than 3 shots.

5. Conclusion

We present an in-context-aware image generation framework capable of learning from a variable number of visual context examples and prompts. Our approach leverages both the visual and text inputs on the same level, resulting in a framework able to learn in a balanced manner from the multimodal inputs. Furthermore, learning from a few context examples showed to be helpful in learning strong visual characteristics, especially if the prompt is not available. Our experiments and user study demonstrate the applicability of our approach across diverse tasks and settings, confirming the improved quality and fidelity over counterpart models.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 2022. 3

[2] Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. Spatext: Spatio-textual representation for controllable image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3

[3] Dina Bashkirova, Jose Lezama, Kihyuk Sohn, Kate Saenko, and Irfan Essa. Masksketch: Unpaired structure-guided masked image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3

[4] Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. Retrieval-augmented diffusion models. In *Advances in Neural Information Processing Systems*, 2022. 2, 3

[5] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2, 3

[6] 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. In *Advances in Neural Information Processing Systems*, 2020. 2, 3

[7] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*, 2022. 2, 3

[8] Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. *arXiv preprint arXiv:2304.00186*, 2023. 3

[9] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 2

[10] Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, Matthew Yu, Abhishek Kadian, Filip Radenovic, Dhruv Mahajan, Kunpeng Li, Yue Zhao, Vladan Petrovic, Mitesh Kumar Singh, Simran Motwani, Yi Wen, Yiwen Song, Roshan Sumbaly, Vignesh Ramanathan, Zijian He, Peter Vajda, and Devi Parikh. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*, 2023. 2

[11] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *European Conference on Computer Vision*, 2022. 3

[12] 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, 3

[13] Vidit Goel, Elia Peruzzo, Yifan Jiang, Dejia Xu, Nicu Sebe, Trevor Darrell, Zhangyang Wang, and Humphrey Shi. Pair-diffusion: Object-level image editing with structure-and-appearance paired diffusion models. *arXiv preprint arXiv:2303.17546*, 2023. 3

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016. 4

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

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANS trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017. 6

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33, 2020. 2

[18] Qingqing Huang, Daniel S Park, Tao Wang, Timo I Denk, Andy Ly, Nanxin Chen, Zhengdong Zhang, Zhishuai Zhang, Jiahui Yu, Christian Frank, et al. Noise2music: Text-conditioned music generation with diffusion models. *arXiv preprint arXiv:2302.03917*, 2023. 2

[19] Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. Grounding language models to images for multimodal inputs and outputs. In *Proceedings of the International Conference on Machine Learning*, 2023. 3

[20] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35, 2022. 3

[21] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 3

[22] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3

[23] Zejian Li, Jingyu Wu, Immanuel Koh, Yongchuan Tang, and Lingyun Sun. Image synthesis from layout with locality-aware mask adaption. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 3

[24] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022. 2

[25] 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. *arXiv preprint arXiv:2108.01073*, 2022. 3

[26] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3

[27] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 2022. 2

[28] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2

[29] Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, et al. UniControl: A unified diffusion model for controllable visual generation in the wild. *arXiv preprint arXiv:2305.11147*, 2023. 3

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, 2021. 3, 5

[31] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Proceedings of the International Conference on Machine Learning*. PMLR, 2021. 2

[32] 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

[33] 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*, 2022. 2, 3, 4

[34] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015. 3

[35] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3

[36] 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*, 2022. 2, 3

[37] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: Learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics*, 2016. 5

[38] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the International Conference on Machine Learning*, 2015. 2

[39] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 2, 5

[40] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2

[41] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 2021. 3

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 4

[43] Zhendong Wang, Yifan Jiang, Yadong Lu, Yelong Shen, Pengcheng He, Weizhu Chen, Zhangyang Wang, and Mingyuan Zhou. In-context learning unlocked for diffusion models. *arXiv preprint arXiv:2305.01115*, 2023. 1, 2, 3, 4, 5, 6, 7, 8

[44] Zhendong Wang, Yifan Jiang, Huangjie Zheng, Peihao Wang, Pengcheng He, Zhangyang Wang, Weizhu Chen, and Mingyuan Zhou. Patch diffusion: Faster and more data-efficient training of diffusion models. *arXiv preprint arXiv:2304.12526*, 2023. 2

[45] 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, 3

[46] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. 2, 3

[47] Xingqian Xu, Jiayi Guo, Zhangyang Wang, Gao Huang, Irfan Essa, and Humphrey Shi. Prompt-free diffusion: Taking" text" out of text-to-image diffusion models. *arXiv preprint arXiv:2305.16223*, 2023. 3

[48] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 2, 3, 4, 5, 7, 8

[49] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 2