Context Diffusion:
In-Context Aware Image Generation

Ivona Najdenkoska[1,2]*   Animesh Sinha[1]   Abhimanyu Dubey[1]
Dhruv Mahajan[1] Vignesh Ramanathan[1] Filip Radenovic[1]

[1]Meta GenAI   [2]University of Amsterdam

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 context fidelity of the generated images deteriorate when the prompt is not present, demonstrating that these models cannot truly learn from the visual context. To address this, we propose a novel framework that separates the encoding of the visual context and the preservation of the desired image layout. This results in the ability to learn from the visual context and prompts, but also from either of them. Furthermore, we enable our model to handle few-shot settings, to effectively address diverse in-context learning scenarios. Our experiments and human evaluation demonstrate that Context Diffusion excels in both in-domain and out-of-domain tasks, resulting in an overall enhancement in image quality and context fidelity compared to counterpart models.

Keywords: Image generation · Diffusion models · In-context learning

1 Introduction

Generative models are witnessing major advances, both in natural language [7,10,30,44,50,55] and media generation [6,11,21,31,36,38,39,43]. Large language models in particular have shown impressive in-context learning capabilities [1,7,45,51]. This represents 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 and domains. However, for generative models in computer vision, learning from context examples remains under-explored.

The closest line of work that explicitly supports a single image pair as a context example for image generation is Prompt Diffusion [48]. It builds on the popular ControlNet [54] model which introduced the idea of controllable diffusion models. Specifically, Prompt Diffusion attempts to learn the visual mapping from a source image to a target context image and applies it to a new query image, by also leveraging a prompt for text-based guidance. However, we empirically

---

* Work done during an internship at Meta GenAI. Correspondence at `i.najdenkoska@uva.nl`.
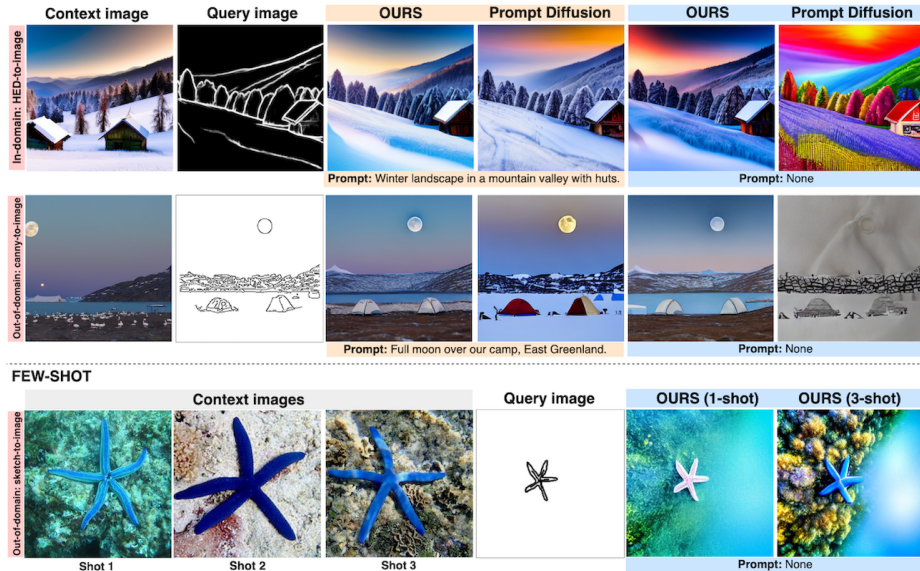
Fig. 1: Illustrating in-context aware image generation with Context Diffusion.
Top row: HED-to-image as an in-domain task; **Middle row**: canny-to-image as an
out-of-domain task. Our model enables learning from context with and without prompts.
The counterpart model, Prompt Diffusion [48] cannot leverage 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.

observed that this model struggles to leverage the context example 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 pair shows specific styles, they
cannot be leveraged 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 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 solely target
examples as visual 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 layout through edges, depth,
segmentation maps, etc. On the other hand, the context examples provide hints

for finer details such as styles, textures, color palettes, and object appearances desired in the output image.

Furthermore, it is important to note the difficulty in controlling both aspects of the output image solely through the control mechanisms [48, 54]. The "control" part of the model is very effective in capturing high-level structure and layouts. However, fine-grained details are better captured through the conditioning mechanism, as seen in textual inversion [14], grounded text-to-image generation [25], and retrieval-augmented image generation [5, 9]. To that end, we inject the visual information from the context into the network as text conditioning. Different from prior works, we aggregate the sequence of all visual embeddings extracted from the context images and place them alongside the text embeddings in the cross-attention layers of the diffusion model. The ability to aggregate a variable number of context images enables the model to handle few-shot scenarios. Moreover, these modified cross-attention layers allow a stronger reliance on the visual context which is especially apparent without textual guidance. The layout of the query image is preserved by passing it as a control signal to the network in a similar manner as ControlNet [54].

At inference time, we use a query image to define the target layout 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 the structure and layout from the query image and transfer visual signals from the context even when the text prompt is not provided. Moreover, to properly study in-context learning abilities, we experiment with unseen, out-of-domain tasks, such as handling sketches, image editing, and more. Furthermore, for such tasks, using multiple images in a few-shot manner yields stronger 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 leverage visual context examples to control the fine-grained details of the output image, alongside a query image that defines the layout 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 prior work.

2   Related Work

Diffusion-based Image Generation. Recent advancements in diffusion models, first introduced in [41] have exhibited huge success in text-to-image generation tasks [11, 20, 34, 35, 39]. Enhancements have been achieved through various

training [11, 36, 39] and sampling [27, 42, 49] techniques. For instance, DALLE-2 [34] proposed an architecture encompassing several stages, by encoding text with CLIP [33] language encoder and decoding images from the encoded text embeddings, followed by Imagen [39] which showed that up-scaling the text encoder largely improves the text fidelity. Furthermore, the Latent Diffusion Model (LDM) [36] 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. Besides the text prompts, adding more control to the image generation process helps overall customization and task-specific image generation. Recent text-conditioned models focus on adjusting models by task-specific fine-tuning [8, 14, 15, 38, 52], injecting conditioning maps, like segmentation maps, sketches or key-points [2, 4, 12, 13, 26, 32, 54, 56], or exploring editing abilities [6, 16, 18, 28, 29]. 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 [25] inject grounding information, such as bounding boxes or edge maps, into new trainable layers via a gated cross-attention mechanism. ControlNet [54], as a recent state-of-the art in controllable image generation presents a general framework for adding spatial conditioning controls. UniControl [32] extends ControlNet by unifying various image map conditions into a single image generation framework. Other works, such as Re-Imagen [9] and RDM [5], 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 learning from in-context 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. Although in-context learning is vastly explored both in language-only [7, 23, 50, 51] and visual-language models [1, 22, 24, 45], its application is lagging behind in image generation. Bar *et al.* [3] investigate visual prompting for image generation, followed by Painter [47] which incorporates more tasks to construct a generalist model. Regarding in-context abilities in diffusion models, Prompt Diffusion [48] presents such a framework that extends the control capability of ControlNet [54] 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.

Another work tackling image generation with visual examples is Prompt-Free Diffusion [53]. It focuses only on having an image as a context *i.e.* 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 $z_{t-1}, \ldots, z_0$, 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, defined as follows:

$$\min_\theta \mathcal{L} = \mathrm{E}_{z,\epsilon \sim \mathcal{N}(0,1),t} \big[\parallel \epsilon - f_\theta(z_t, t, \mathbf{c}) \parallel_2^2 \big], \tag{1}$$

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 [54], for encoding the layout of the desired output via a query image as visual control. Note that 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 text embeddings.

3.2   Context Diffusion Architecture

The model f_θ is essentially a UNet architecture [37], with an encoder, a middle block, and a skip-connected decoder. These modules denoted as LDM encoder, mid, and decoder in Figure 2, are built out of standard ResNet [17] and Transformer blocks [46] which contain several cross-attention and self-attention mechanisms. The core of conditional-diffusion models is encoding the conditional information [36], based on which z_t is generated at a given time step t. We differentiate two types of such conditional information: the visual context $\mathbf{V}$
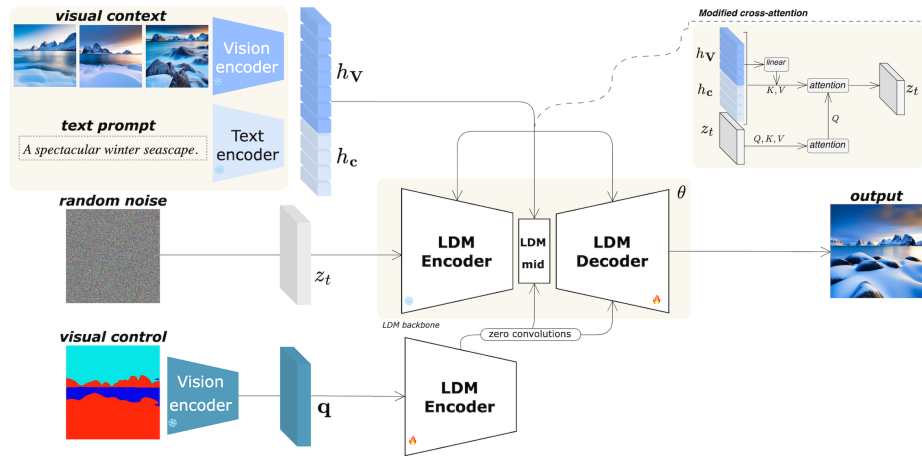
Fig. 2: Architecture of Context Diffusion. It consists of several modules: vision and text encoders for encoding the text prompt and visual context respectively, an LDM backbone for handling the image generation process, and an additional LDM encoder for processing the query image as a visual control. Note that here we show three visual context examples, however, the model is trained using a variable number of such examples.

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 denotes the number of images. Additionally, we consider a visual control image, *i.e.*, the query image that serves to define the layout of the output denoted as $\mathbf{q}$.

Prompt encoding. To perform the encoding of the textual prompt $\mathbf{c}$ we use a pretrained language encoder f_{text} with fixed parameters to obtain the embeddings. Particularly, we obtain $\mathbf{h}^{\mathbf{c}} = \{h_0^{\mathbf{c}}, \dots, h_{N^{\mathbf{c}}}^{\mathbf{c}}\} = f_{\text{text}}(\mathbf{c})$, 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 $\mathbf{h}^{\mathbf{V}} = \sum_{i=1}^{k} \mathbf{h}^{\mathbf{v}_i}$. Then, 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:

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

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 concatenate them to obtain $[\mathbf{h^c}, \mathbf{h^V}]$, illustrated in the bottom left corner of Figure 2. Thus the input to the cross-attention block in (2) changes as follows:

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

Visual control encoding. To enable the ingestion of the query image as visual control, we follow ControlNet setup [54]. 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 $\mathbf{y}$:

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

To train with this objective, we use a collection of tasks for joint end-to-end training, similar to [48]. Different from them, we use a visual context sequence consisting of a k-images and an optional 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 a variable number of context images and to apply them along with the text prompt to the query image.

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 [48]. Following their training setup and code implementation, 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 usually related and describe a similar conditioning signal.

Fig. 3: In-domain comparison to Prompt Diffusion [48]: Examples of {HED, segmentation, depth}-to-image as forward tasks and image-to-{HED, segmentation, depth} as reverse tasks, with both visual context and prompt (C+P) given as conditioning information.

At inference time, we use the test partition of the dataset to test the ability of the model 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 utilize hand-drawn sketches from the Sketchy dataset [40], where the sketch is the query image, and the real images are the visual 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 [54].

Implementation Details. The backbone of our model follows a vanilla ControlNet architecture, initialized in the same way as [48]. We train the 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 [33] 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. Following prior works [25, 48, 54], we apply random replacement of the prompts with empty strings for classifier-free guidance. We experimented with different rates and found that 50% is the most optimal one, as also reported by ControlNet [54]. At inference time, we apply DDIM [42] 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.

Human Evaluation Setup. To better quantify the performance, we perform a human evaluation to compare our model to Prompt Diffusion [48]. A total of 10 in-house annotators participated in the study, annotating 240 randomly chosen
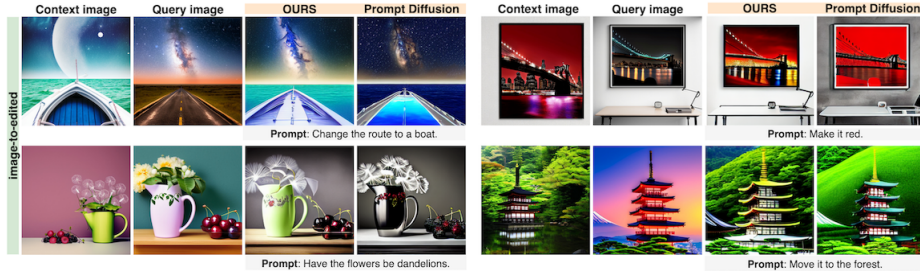
Fig. 4: Out-of-domain comparison to Prompt Diffusion [48]: Image editing, with visual context and prompt (C+P) as conditioning information.

Fig. 5: Out-of-domain comparison to Prompt Diffusion [48]: {sketch, normal map, scribble, canny edge}-to-image tasks. Visual context and prompt (C+P) are given as conditioning information.

samples from the test partition. For each test sample, we present two images - one generated by our Context Diffusion and another by Prompt Diffusion, randomly annotated as A and B, alongside the given visual context, query image, and prompt. Each annotator chooses either a preferred image or denotes 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 (C+P), using only visual context (C) and only prompt (P). Considering all these scenarios is highly important since it 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 Evaluation. In addition to the human evaluation, we also use offline automated metrics to further evaluate the performance of Context Diffusion. In particular, we report FID [19] scores for map-to-image tasks and RMSE scores for the image-to-map tasks, by taking the average across three random seeds. Note that we use 5000 randomly chosen test samples per setting for each task to generate output images both with 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 [48]. Prompt Diffusion expects a source-target

Table 1: Human evaluation comparison to Prompt Diffusion (PD) [48]: 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.

	In-domain				Out-of-domain			
	C+P	C	P	*avg*	C+P	C	P	*avg*
PD [48]	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**

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 4.2 Ablation study. It is important to notice that in all comparisons we follow the source-target format of the input to generate images with Prompt Diffusion, 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 at inference time. We also present the generated results of our model on *few-shot* setup when several visual examples are given as input. Prompt Diffusion does not support the few-shot setup. Finally, we present the ablations for the key components of our proposed model.

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 four columns depict the forward tasks, while the last four columns depict the reverse tasks. It can be observed that our model can 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 the human evaluation presented in Table 1 (In-domain (C+P) column), as well as in offline metrics comparison presented in Table 2 (C+P columns), where we obtain satisfactory performance improvement (36.3% *vs.* 28.5% win-rate) over Prompt Diffusion.
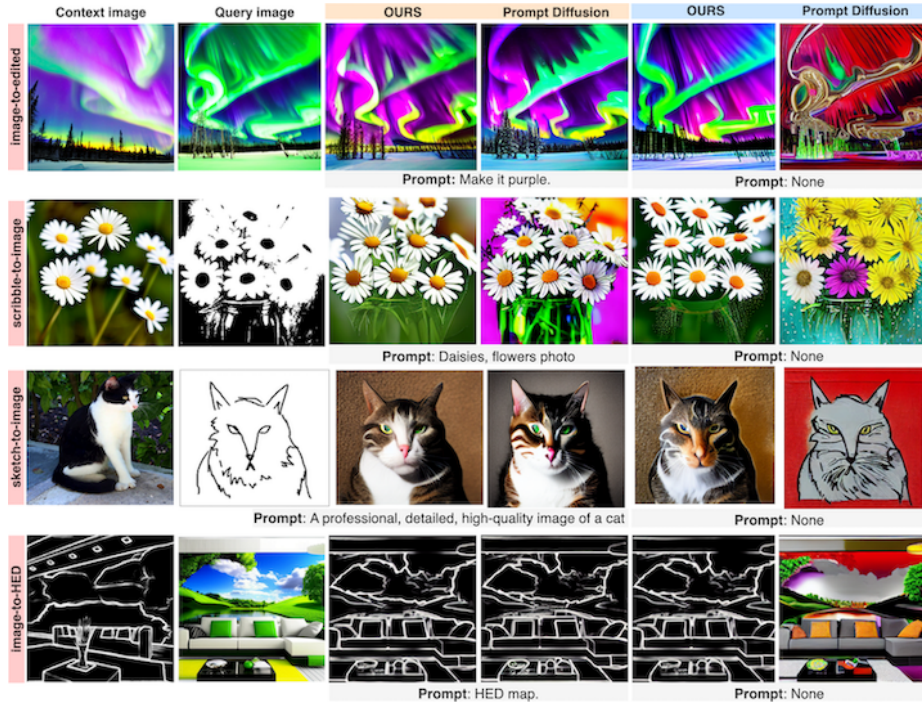
Fig. 6: Conditioning comparison with Prompt Diffusion [48]: Using visual context and prompt (C+P) and visual context-only (C) as conditioning, on in-domain (image-to-HED) and out-of-domain (editing, scribble-to-image, sketch-to-image) tasks.

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 the 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 [48]. 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 human evaluation 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.

Conditioning variations at inference

Using only visual context. To better understand the effect of visual context examples on the model's performance, we analyze the outputs when the text
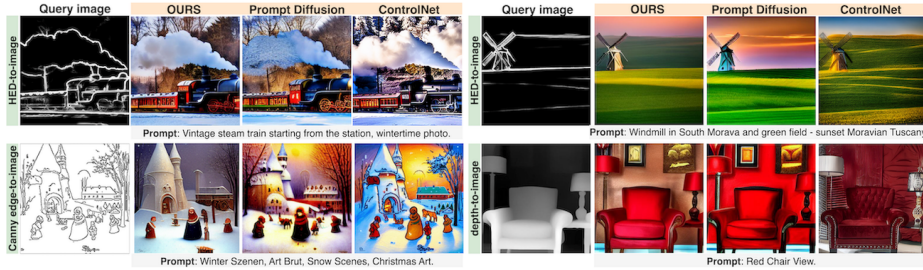
Fig. 7: Zero-shot comparison to ControlNet [54] and Prompt Diffusion [48]: Using prompt-only (P) as conditioning information.

Table 2: Offline comparison to Prompt Diffusion (PD) [48] using 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.

| | FID (map-to-img) ↓ | | | | | | | | |
| | HED-to-img | | | seg-to-img | | | depth-to-img | | |
	C+P	C	P	C+P	C	P	C+P	C	P
PD [48]	12.8	22.5	15.1	16.7	25.1	**17.2**	15.9	27.0	18.1
Ours	**12.3**	**17.7**	**14.8**	**13.4**	**19.0**	18.5	**12.9**	**18.5**	**17.5**

| | RMSE (img-to-map) ↓ | | | | | | | | |
| | img-to-HED | | | img-to-seg | | | img-to-depth | | |
	C+P	C	P	C+P	C	P	C+P	C	P
PD [48]	0.15	0.33	**0.15**	0.32	0.41	0.32	**0.14**	0.34	0.14
Ours	**0.11**	**0.11**	0.16	**0.29**	**0.28**	**0.30**	**0.14**	**0.13**	**0.13**

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 [48] is unable to learn from the visual examples, indicating that it relies solely on the text caption as conditional information. We include this setting in the human evaluation 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 2 ((C) columns) for in-domain tasks, further strengthening the observations that our model can truly leverage styles, color palettes, and object appearances in 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. This essentially yields zero-shot settings, boiling down to how ControlNet [54] 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
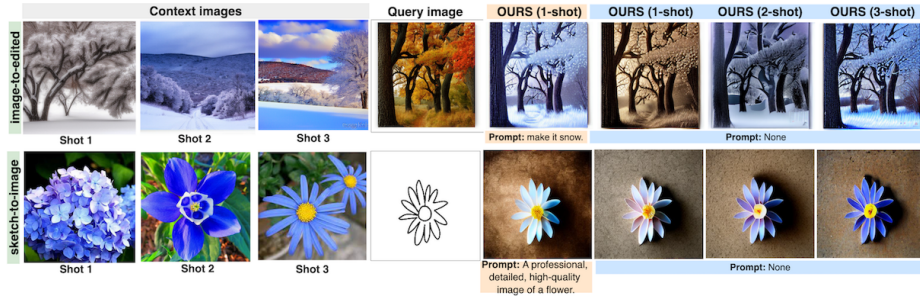
Fig. 8: Few-shot examples: Comparison between out-of-domain tasks (editing and sketch) using one context example with a text prompt, and one, two, and three shots of context examples with no text prompt. Our model can leverage multiple visual examples to handle scenarios when the text prompt is not present.

Figure 7 we show representative examples of this setting, comparing our model to ControlNet and Prompt Diffusion. It can be seen that our model can generate more realistic images compared to ControlNet and Prompt Diffusion. Similar to before, we also include this setting in the user study, reported in Table 1 ((P) columns). We observe much better performance on the out-of-domain (49.8% *vs.* 25.9%) tasks compared to Prompt Diffusion, and a slight decrease on in-domain (29.6% *vs.* 30.4%) tasks. This supports our observations that Prompt Diffusion 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 2 ((P) columns), again observing better overall performance across different tasks.
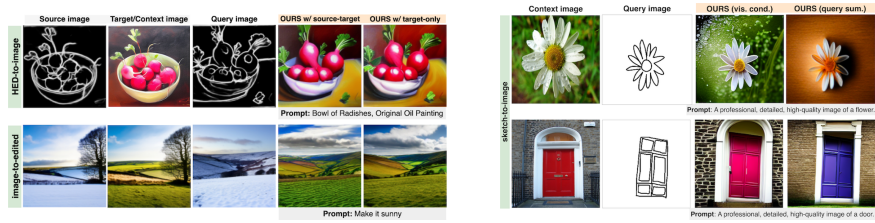
Few-shot visual context examples Context Diffusion is designed to handle multiple context examples, enabling few-shot scenarios. Using one context example proved to be enough for in-domain tasks, as seen in Figure 3, even without using a text prompt as seen in Figure 6. Therefore in the few-shot experiments, we focus on the out-of-domain tasks, such as editing and sketch-to-image. In particular, we augment the visual context with additional images, depicting similar objects, scenes, or desired color palettes. 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

Table 3: Human evaluation 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.

	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**

can be seen from Figure 8, multiple demonstrations of images help to strengthen the target visual representation, especially when the prompt is not present. We also quantify the performance by conducting a human evaluation for the few-shot settings, presented in Table 3. We compare our model when using one context

(a) Source-target vs target-only as context. Having a source image paired with the target does not influence the output image, since the visual information is provided by the target only.

(b) Benefit of visual conditioning. Ingesting the context examples as visual conditioning helps in leveraging the visual cues, unlike the summation of the context to the query image.

Fig. 9: **Ablations.** We ablate the key components of our framework, namely (a) Using source-target vs target-only as context, (b) Benefit of visual conditioning.

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. 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.

Ablations

Source-target vs target-only as context. We analyze the performance when training by using source-target image pairs as context examples (same as Prompt Diffusion [48]). Figure 9a shows that there is no benefit when the source image is paired with the target image during the training. The visual information needed for generating the image is entirely contained in the target *i.e.* context image or the prompt, while the query image controls the layout.

Benefit of visual conditioning. Our model employs a visual conditioning pipeline to ingest the context examples similarly to the text prompts. To better understand the benefit of such conditioning, we compare it to the same paradigm used in Prompt Diffusion [48]. Specifically, we use ConvNets to encode the context examples and then we sum them to the query image. The entire representation is then fed into the ControlNet module. As we can see in Figure 9b, training in this manner prevents the model from actually leveraging the context.

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 as conditioning information, 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 human evaluation study demonstrate the applicability of our approach across diverse tasks and settings, confirming the improved quality and context fidelity over prior work.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems (2022)
2. Avrahami, O., Hayes, T., Gafni, O., Gupta, S., Taigman, Y., Parikh, D., Lischinski, D., Fried, O., Yin, X.: Spatext: Spatio-textual representation for controllable image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
3. Bar, A., Gandelsman, Y., Darrell, T., Globerson, A., Efros, A.A.: Visual prompting via image inpainting. In: Oh, A.H., Agarwal, A., Belgrave, D., Cho, K. (eds.) Advances in Neural Information Processing Systems (2022)
4. Bashkirova, D., Lezama, J., Sohn, K., Saenko, K., Essa, I.: Masksketch: Unpaired structure-guided masked image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
5. Blattmann, A., Rombach, R., Oktay, K., Müller, J., Ommer, B.: Retrieval-augmented diffusion models. In: Advances in Neural Information Processing Systems (2022)
6. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
7. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems (2020)
8. Chen, W., Hu, H., Li, Y., Ruiz, N., Jia, X., Chang, M.W., Cohen, W.W.: Subject-driven text-to-image generation via apprenticeship learning. arXiv preprint arXiv:2304.00186 (2023)
9. Chen, W., Hu, H., Saharia, C., Cohen, W.W.: Re-imagen: Retrieval-augmented text-to-image generator. arXiv preprint arXiv:2209.14491 (2022)
10. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al.: Palm: Scaling language modeling with pathways. arXiv preprint arXiv:2204.02311 (2022)
11. Dai, X., Hou, J., Ma, C.Y., Tsai, S., Wang, J., Wang, R., Zhang, P., Vandenhende, S., Wang, X., Dubey, A., Yu, M., Kadian, A., Radenovic, F., Mahajan, D., Li, K., Zhao, Y., Petrovic, V., Singh, M.K., Motwani, S., Wen, Y., Song, Y., Sumbaly, R., Ramanathan, V., He, Z., Vajda, P., Parikh, D.: Emu: Enhancing image generation models using photogenic needles in a haystack. arXiv preprint arXiv:2309.15807 (2023)
12. Fan, W.C., Chen, Y.C., Chen, D., Cheng, Y., Yuan, L., Wang, Y.C.F.: Frido: Feature pyramid diffusion for complex scene image synthesis. In: Proceedings of the AAAI conference on artificial intelligence. vol. 37, pp. 579–587 (2023)
13. Gafni, O., Polyak, A., Ashual, O., Sheynin, S., Parikh, D., Taigman, Y.: Make-a-scene: Scene-based text-to-image generation with human priors. In: European Conference on Computer Vision (2022)
14. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)
15. Gal, R., Arar, M., Atzmon, Y., Bermano, A.H., Chechik, G., Cohen-Or, D.: Encoder-based domain tuning for fast personalization of text-to-image models. ACM Transactions on Graphics (TOG) (2023)

16. Goel, V., Peruzzo, E., Jiang, Y., Xu, D., Sebe, N., Darrell, T., Wang, Z., Shi, H.: Pair-diffusion: Object-level image editing with structure-and-appearance paired diffusion models. arXiv preprint arXiv:2303.17546 (2023)

17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2016)

18. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)

19. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANS trained by a two time-scale update rule converge to a local nash equilibrium. Advances in Neural Information Processing Systems **30** (2017)

20. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33** (2020)

21. Huang, Q., Park, D.S., Wang, T., Denk, T.I., Ly, A., Chen, N., Zhang, Z., Zhang, Z., Yu, J., Frank, C., et al.: Noise2music: Text-conditioned music generation with diffusion models. arXiv preprint arXiv:2302.03917 (2023)

22. Koh, J.Y., Salakhutdinov, R., Fried, D.: Grounding language models to images for multimodal inputs and outputs. In: Proceedings of the International Conference on Machine Learning (2023)

23. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large language models are zero-shot reasoners. Advances in Neural Information Processing Systems **35** (2022)

24. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)

25. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)

26. Li, Z., Wu, J., Koh, I., Tang, Y., Sun, L.: Image synthesis from layout with locality-aware mask adaption. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021). `https://doi.org/10.1109/ICCV48922.2021.01356`

27. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. arXiv preprint arXiv:2211.01095 (2022)

28. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. arXiv preprint arXiv:2108.01073 (2022)

29. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: 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)

30. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems **35** (2022)

31. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022)

32. Qin, C., Zhang, S., Yu, N., Feng, Y., Yang, X., Zhou, Y., Wang, H., Niebles, J.C., Xiong, C., Savarese, S., et al.: UniControl: A unified diffusion model for controllable visual generation in the wild. arXiv preprint arXiv:2305.11147 (2023)

33. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proceedings of the International Conference on Machine Learning (2021)

34. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)

35. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: Proceedings of the International Conference on Machine Learning. PMLR (2021)

36. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)

37. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer (2015)

38. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: 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)

39. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems (2022)

40. Sangkloy, P., Burnell, N., Ham, C., Hays, J.: The sketchy database: Learning to retrieve badly drawn bunnies. ACM Transactions on Graphics (2016)

41. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: Proceedings of the International Conference on Machine Learning (2015)

42. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2021)

43. Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)

44. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)

45. Tsimpoukelli, M., Menick, J.L., Cabi, S., Eslami, S., Vinyals, O., Hill, F.: Multimodal few-shot learning with frozen language models. Advances in Neural Information Processing Systems (2021)

46. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems (2017)

47. Wang, X., Wang, W., Cao, Y., Shen, C., Huang, T.: Images speak in images: A generalist painter for in-context visual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6830–6839 (2023)

48. Wang, Z., Jiang, Y., Lu, Y., Shen, Y., He, P., Chen, W., Wang, Z., Zhou, M.: In-context learning unlocked for diffusion models. arXiv preprint arXiv:2305.01115 (2023)

49. Wang, Z., Jiang, Y., Zheng, H., Wang, P., He, P., Wang, Z., Chen, W., Zhou, M.: Patch diffusion: Faster and more data-efficient training of diffusion models. arXiv preprint arXiv:2304.12526 (2023)
50. Wei, J., Bosma, M., Zhao, V.Y., Guu, K., Yu, A.W., Lester, B., Du, N., Dai, A.M., Le, Q.V.: Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652 (2021)
51. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al.: Emergent abilities of large language models. Transactions on Machine Learning Research (2022)
52. Wei, Y., Zhang, Y., Ji, Z., Bai, J., Zhang, L., Zuo, W.: Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15943–15953 (2023)
53. Xu, X., Guo, J., Wang, Z., Huang, G., Essa, I., Shi, H.: Prompt-free diffusion: Taking" text" out of text-to-image diffusion models. arXiv preprint arXiv:2305.16223 (2023)
54. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
55. Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., et al.: Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068 (2022)
56. Zhao, S., Chen, D., Chen, Y.C., Bao, J., Hao, S., Yuan, L., Wong, K.Y.K.: Uni-ControlNet: All-in-one control to text-to-image diffusion models. Advances in Neural Information Processing Systems (2024)