# Beyond Textual Constraints: Learning Novel Diffusion Conditions with Fewer Examples

Yuyang Yu [1*] Bangzhen Liu [1*] Chenxi Zheng [1]
Xuemiao Xu[1,2,3,4†] Huaidong Zhang [1†] Shengfeng He [5]
[1]South China University of Technology [2]State Key Laboratory of Subtropical Building Science
[3]Guangdong Provincial Key Lab of Computational Intelligence and Cyberspace Information
[4]Ministry of Education Key Laboratory of Big Data and Intelligent Robot [5]Singapore Management University

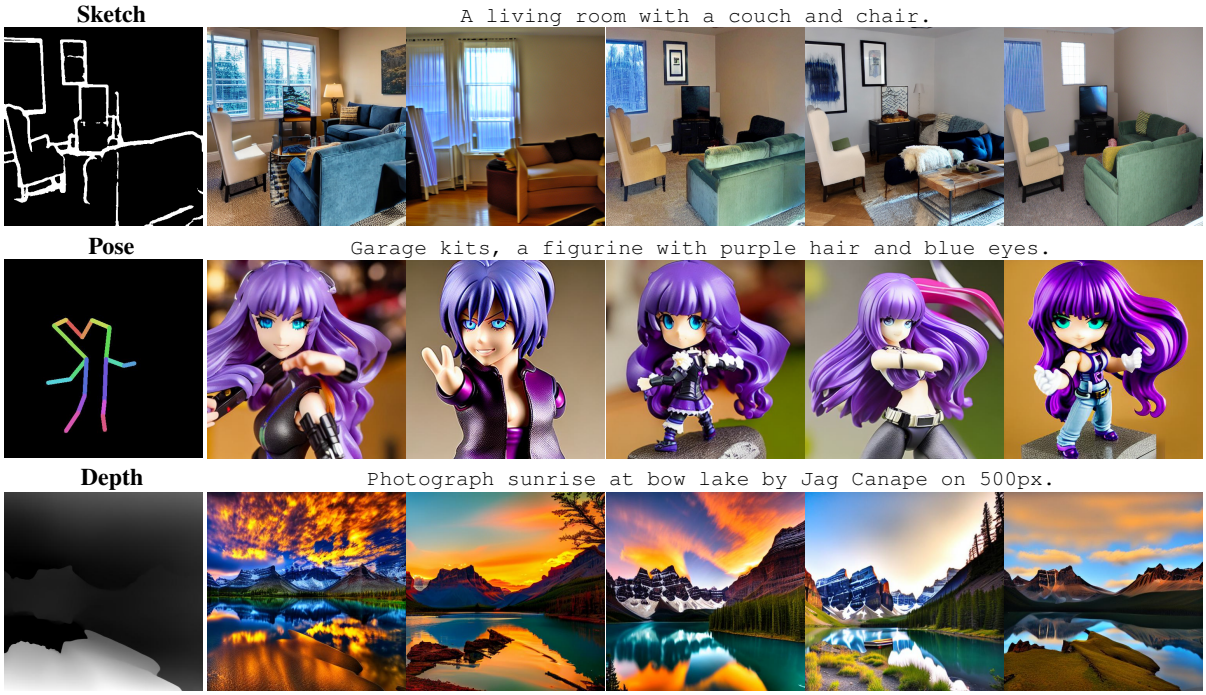



Figure 1. We explore a novel aspect of learning diffusion conditions, requiring only a magnitude of a thousand times fewer examples (only 100 vs. 100k) compared to existing methods like ControlNet [44] and T2IAdapter [26]. The "-100" suffix in our model names indicates training with just 100 text-image-condition pairs. Our method achieves both structural consistency and high-quality generation with these limited samples, delivering performance comparable to the fully trained models of our competitors.

## Abstract

*In this paper, we delve into a novel aspect of learning novel diffusion conditions with datasets an order of magnitude smaller. The rationale behind our approach is the elimination of textual constraints during the few-shot learning process. To that end, we implement two optimization strategies. The first, prompt-free conditional learning, utilizes a prompt-free encoder derived from a pre-trained Stable Diffusion model. This strategy is designed to adapt new conditions to the diffusion process by minimizing the textual-visual correlation, thereby ensuring a more precise alignment between the generated content and the specified conditions. The second strategy entails condition-specific negative rectification, which addresses the inconsistencies typically brought about by Classifier-free guidance in few-shot training contexts. Our extensive experiments across a variety of condition modalities demonstrate the effectiveness and efficiency of our framework, yielding results comparable to those obtained with datasets a thousand times larger. Our codes are available at https://github.com/Yuyan9Yu/BeyondTextConstraint.*

*The first two authors contributed equally.
†Corresponding authors (xuemx@scut.edu.cn, huaidongz@scut.edu.cn).

Segmentation (a) w. text (b) w/o. text

Figure 2. Segmentation-conditioned text-to-image generation of ControlNet-100 w. and w/o. text condition. Incorporating text constraints would lead to structurally inconsistent regions (bounded by red boxes), when only limited training exemplars are available.

## 1. Introduction

Large-scale generative models [9, 18, 19, 33], notably the Stable Diffusion (SD) series [33], have significantly advanced image synthesis. These models produce high-fidelity images with visually striking content from concise textual prompts. However, despite the versatility of textual descriptions in directing the visual elements of generated images, they frequently lack precision in conveying intricate details such as spatial layouts, poses, shapes, and forms, when relying solely on text prompts.

To enhance user control over the content generated by incorporating additional exemplar images, ControlNet [44] and T2IAdapter [26] have been developed to augment pretrained SD models with controllable adapters. These adapters introduce novel conditioning factors without finetuning the original SD models, enriching the capability of them. While these encoders are effective in adapting to new conditions, they require extensive fine-tuning with large datasets (at least tens of thousands of examples) for each unique modality of conditioning. The necessity for substantial data collection and meticulous annotation imposes significant limitations in terms of both financial and time resources. This requirement often renders it infeasible for users to acquire sufficient data for the adaptation process, thus limiting practical applicability in real-world scenarios.

Given the aforementioned challenges, we are prompted to explore the question: 'Is it possible for diffusion models to effectively learn novel diffusion conditions with extremely limited training data of a scale manageable by ordinary users?' This paper investigates a pertinent issue in the field: adapting pre-trained Text-to-Image generative models to new condition modalities with a scarce number of exemplars. A naive strategy, such as directly fine-tuning the model on a constrained dataset, might lead to discrepancies with the intended condition [16]. This issue is notably apparent in the misalignment of spatial structures, including sketches, semantic maps, or poses.

This phenomenon can be attributed to the complex interplay between text prompts and image features, especially through the cross-attention mechanisms in the SD model. The data distributions across different modalities exhibit significant domain gaps. To bridge these gaps, the original SD model employs multiple cross-attention layers, aiming to create a unified and flexible feature space for text-guided image generation. However, when dealing with a novel condition characterized by sparse training data, the model struggles to reconfigure the feature space to align coherently with the new condition and associated text. Consequently, the pronounced bias from text prompts leads to inadequate learning from the novel condition, potentially resulting in generation outcomes that are structurally misaligned with the intended condition, as compared in Fig. 2a and Fig. 2b.

Depth Segmentation

Figure 3. Conditioned text-to-image generation of our method after introducing prompt-free conditional learning. The text prompts for depth and segmentation are "Dragon and phoenix in fantasy world art.", and "A piece of chocolate cake on a plate.", respectively. While the generated images are more structurally consistent with the given exemplar, however, part of the contents (bounded by red rectangles) in the aligned region disagree with the given text.

Motivated by our earlier analysis, our objective is to reduce the biased impact of text prompts in the learning of structural conditions within a few-shot learning context. To this end, we introduce a two-stage optimization framework comprising *prompt-free conditional learning* and *condition-specific negative rectification*. The first stage involves a prompt-free encoder, initially adapted from a pre-trained SD model, to utilize the established text-image feature space. This step involves progressively diminishing the influence of text descriptions by deactivating the text connection within the prompt-free encoder. This approach, combined with a null-text fine-tuning strategy, enables the prompt-free encoder to learn novel conditions with minimal textual influence, thereby achieving improved structural alignment between generated content and the given condition. The second stage of our approach is driven by the observation that utilizing Classifier-Free Guidance [13] (CFG) in the diffusion process can occasionally result in inconsistent content during few-shot training (see Fig. 3). This inconsistency primarily stems from the application of a universal negative prompt across all training samples. While this prompt offers directional guidance, it is tailored for full-scale data and proves to be less effective, or even counterproductive, in the context of few-shot training. The universal negative prompt, designed for larger datasets, inadequately guides the learning process when only a limited number of examples are available. To

address this, we propose a condition-specific negative rectification method. This involves using a lightweight encoder to adjust the negative prompt based on the structurally conditional embedding of the exemplar image. Consequently, the CFG process can offer more precise structural guidance, enhancing the overall quality of generated content. Owing to the efficiency of our few-shot learning framework, it is possible to adapt to new conditions using an order of magnitude fewer data points compared to ControlNet [44] and T2IAdapter [26], while still achieving comparable results (see Fig. 1e).

Our contributions can be summarized as follows:

- We present the first few-shot novel diffusion condition learning framework, capable of adjusting to new conditions with significantly fewer training samples.
- We propose the prompt-free conditional learning to eliminate the text constraints during the learning of novel conditions, and further improve the generation quality by a condition-specific negative rectification.
- We conduct extensive generative experiments on five modalities of conditions, demonstrating the effectiveness and efficiency of the proposed framework.

## 2. Related Work

**Diffusion Models for Image Generation.** Diffusion model [38] has swiftly demonstrated successful applications in the realm of image generation [3, 7, 12, 15, 23, 29, 33, 39, 45, 47, 48]. The remarkable text encoding capabilities of pretrained language models, such as CLIP [31] and BERT [20], have facilitated the diffusion model's widespread adoption in text-to-image generation tasks, yielding exceptional performance. VQ Diffusion [10] executes text-to-image generation through a Mask-and-Replace diffusion strategy in latent image space [9], utilizing text input encoded with CLIP. GLIDE [28] conducts text-to-image generation by replacing the original class label in CFG with textual information. Imagen [35] follows a similar approach to GLIDE [28] but leverages a pretrained language model with enhanced text encoding capabilities. Stable Diffusion [33] revolutionary advances the realm of text-to-image generation by training on extensive datasets LAION [36]. While these text-to-image methods can achieve high-quality generation, it is crucial to note that the text descriptions may not offer an intuitively clear structural indication. In this context, PITI [41] employs structural conditions by reducing the feature distance between structural conditions and text descriptions. Voynov *et al*. [40] introduce a latent edge predictor to align the intermediate features with the given sketch. However, these methods are focused on single condition generation, without incorporating the conditions of different modalities, thereby limiting their fine-grained control ability.

**Controllable Diffusion Models.** Recent researches [2, 6, 14, 16, 21, 26, 30, 37, 43, 44, 46] have emerged to enhance the controllability of text-to-image models. Specifically, ControlNet [44] employs zero-initialized layers [27], while T2I-Adapter [26] learns a lightweight adapter on the frozen, pretrained T2I diffusion model to adapting new conditions. HumanSD [16] employs a novel heatmap-guided denoising loss for skeleton-guided controllable human picture generation. Prompt diffusion [43] introduces a visual language prompt to facilitate contextual learning in diffusion-based generative models. Besides, DiffBlender [21], Uni-ControlNet [46], and UniControl [30] are acquiring the ability to learn a multi-condition controllable diffusion model. Composer [14] proposed a robust diffusion model to improve the controllability over both single and multiple conditions. While these methods advance the realm of controllable T2I diffusion models, the training process demands a substantial amount of condition-image data pairs. Multidiffusion [2] and ZestGuide [6] propose to use segmentation maps for spatial guidance generation without training, but they cannot be extended to other spatial conditions. The necessity of considerable time and effort for the collection and processing of training samples, poses an urgent need for fast and accurate novel condition learning from limited samples.

## 3. Methodology

### 3.1. Preliminaries

**Stable Diffusion (SD).** Stable Diffusion [33] is a large text-to-image model which is composed of an autoencoder and a UNet [34] denoiser. The autoencoder maps an input image $x_0$ to a latent $z_0$ and reconstructs it back to the image. The UNet [34] denoiser, which is parameterized with $\theta$, is responsible for denoising a sampled normal noise map to a meaningful latent conditioned on the text conditions. The objective for optimizing SD is defined as follows:

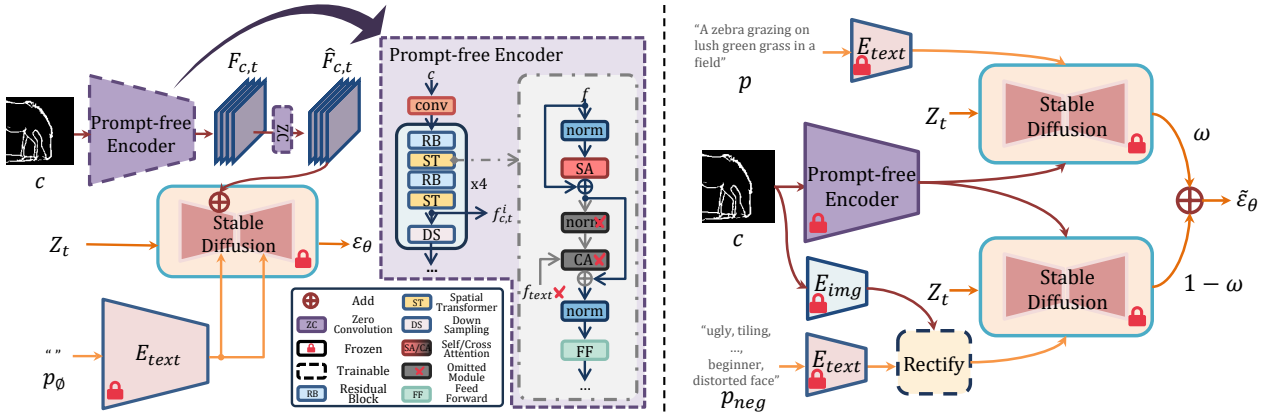
$$L_{sd} = \mathbb{E}_{z,\varepsilon,t}[\|\varepsilon - \varepsilon_\theta(\sqrt{\bar{\alpha}_t} z_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon, c, t)\|^2], \quad (1)$$

where $t$ is the current diffusion time-step; $\varepsilon_\theta$ is the noise predicted by the UNet at time-step $t$; $\varepsilon$ is the corresponding ground truth Gaussian noise; $c$ is the embedding of the textual condition generated by CLIP [31]; and $\bar{\alpha}_t$ is a value of a predefined sequence to facilitate the diffusion process.

**Classifier-Free Guidance (CFG).** To improve the quality of text-conditioned image generation, Ho *et al*. [13] introduce the CFG technique, where the noise prediction is also executed unconditionally. The final noise map used for denoising in CFG is obtained by extrapolating between the conditional and unconditional prediction, which is defined as follows:

$$\tilde{\varepsilon}_\theta(z_t, t, c, \varnothing) = w \cdot \varepsilon_\theta(z_t, t, c) + (1-w) \cdot \varepsilon_\theta(z_t, t, \varnothing), \quad (2)$$

where $\varnothing$ denotes the embedding of a null text; and $\omega$ is the guidance scale parameter, which is often set to 7.5. Note




(a) Prompt-free Conditional Learning

(b) Condition-specific Negative Rectification

Figure 4. Our pipeline involves two stages, prompt-free conditional learning and condition-specific negative rectification. For prompt-free conditional learning, we design a prompt-free encoder to encode the condition and finetune the encoder by incorporating null text conditions with encoded conditional features within the frozen SD model. For condition-specific negative rectification, we rectify the negative prompt with the condition features during the CFG process to achieve more precise diffusion guidance.

that in practical scenarios, SD would rather employ a negative prompt than a null text to achieve enhanced generation results.

### 3.2. Overview

Our aim is to address the problem of adapting the pretrained T2I model to novel conditions, which is limited by textual constraints under the scenario of limited training samples. In this section, we introduce a two-stage optimization framework comprising Prompt-free Conditional Learning (PCL) and Condition-specific Negative Rectification (CNR). The overall framework is shown in Fig. 4. Through a null-text fine-tuning strategy, PCL (Sec. 3.3) employs a prompt-free encoder to explicitly mitigate the biased learning of textual constraints during the adaptation of novel conditions. To further reduce the potential inconsistency and improve the generation quality, we proposed CNR (Sec. 3.4), which incorporates a negative rectifier to dynamically adjust the negative prompt based on the structural condition embedding of the exemplar images.

### 3.3. Prompt-free Conditional Learning

To mitigate the biased learning of textual constraints, we design a prompt-free encoder with a null-text fine-tuning strategy, enabling the learning of novel control conditions with a limited number of training samples.

**Prompt-Free Encoder.** As illustrated in Fig. 4a, the architecture of the prompt-free encoder is modified based on the encoder of SD model, which consists of an input hint layer and 4 sequentially connected blocks. Each block is alternately stacked with two residual blocks and two spatial transformers, followed by a down-sampling layer. Such that each block can be applied for capturing conditional features $F_c^i$ in different scales. We prune the text-conditional links hidden in the spatial transformer to remove the effect of text conditions and initialize the prompt-free encoder with the rest of the pretrained SD encoder. This modification enables us to significantly leverage the prior in the SD model during the novel condition learning, meanwhile diminishing the gap between the output features of the prompt-free encoder and the pretrained text-image feature space when dealing with a limited number of training samples.

Starting from a conditional input $c$ with an original resolution of $512 \times 512$ at timestep $t$, the prompt-free encoder first maps the condition into a $64 \times 64$ feature through the input hint block, and then forward it to obtain the multi-scaled conditional features $F_{c,t} = \{f_{c,t}^i, i \in [1,2,3,4]\}$ with resolutions $\{64, 32, 16, 8\}$, respectively. Subsequently, $F_{c,t}$ are further characterized by their corresponding zero convolution layer $zc^i$, resulting in the derivation of the transformed conditional feature $\hat{F}_{c,t} = \{\hat{f}_{c,t}^i, i \in [1,2,3,4]\}$, where $\hat{f}_{c,t}^i = zc^i\left(f_{c,t}^i\right)$.

In a manner similar to T2I-Adapter [26], We then fuse the $\hat{F}_{c,t}$ to the diffusion process by adding them with the corresponding positional features in the UNet encoder of the SD model.

**Null Text Fine-tuning Strategy.** To further mitigate the negative impact of text conditions, we adopted a null text fine-tuning strategy during the prompt-free conditional learning. This strategy uniformly replaces the original input text with null text $p_\varnothing$, thereby progressively diminishing the influence of text descriptions during the network fine-tuning process.

The objective function employed for the prompt-free conditional learning shares the same form as Eq. 1, which is defined as follows:

$$L_{pcl} = \mathbb{E}_{\boldsymbol{z},\boldsymbol{\varepsilon},\boldsymbol{t}}[\|\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}_{\boldsymbol{\theta}}(\sqrt{\bar{\alpha}_t} z_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon, \hat{F}_{c,t}, \varnothing, t)\|^2], \tag{3}$$

where $\varnothing = E_{text}(p_\varnothing)$ represents the embedding of the null

text $p_\varnothing$ proceeded by the CLIP text encoder $E_{text}$.

By employing the prompt-free encoder and the null text strategy, we successfully eliminate the textual influence and inject controllability of novel conditions into the SD model, thereby elevating image quality under a few-shot setting.

### 3.4. Condition-specific Negative Rectification

Prompt-free conditional learning could significantly achieve more structurally aligned generation with the provided condition, however, there still occasionally exists inconsistent contents. This is due to the universal negative prompt designed for large-scale datasets during the CFG process, which is inadequate for guiding the diffusion process when only limited training samples are available. In order to address this problem, inspired by [8, 25], we propose the condition-specific negative rectification to rectify the universal negative prompt in a more exemplar-specific form for accurate guidance.

**Negative Rectifier.** To address content inconsistency arising from CFG with limited training data, we introduce a negative rectifier. As shown in Fig. 4b, the negative rectifier is a lightweight encoder that can dynamically adjust the universal negative prompt based on the embedding of novel structural conditions. This ensures the acquisition of an exclusive, more suitable negative prompt embedding of each input structural condition for generation.

Specifically, the negative rectifier is a fundamental spatial transformer block, consisting of a self-attention module, a cross-attention module, a feedforward layer, and multiple normalization layers. We utilize the image encoder $E_{img}$ of CLIP to handle the new structural condition image $c$ and derive the structural condition embedding $R_{cond}$. Simultaneously, The text encoder $E_{text}$ of CLIP processes the negative prompt $p_{neg}$ for the negative prompt embedding $R_{neg}$. Subsequently, the $R_{neg}$ and $R_{cond}$ are blended in the negative rectifier through cross-attention operation. Finally, we could obtain the modified negative prompt embedding $\hat{R}_{neg}$, which is employed as the text condition input for the SD during the CFG inference. The described process can be formally defined by the following formula:

$$\begin{aligned} R_{cond} &= E_{img}(c), \\ R_{neg} &= E_{text}\left(p_{neg}\right), \\ \hat{R}_{neg} &= E_{nr}\left(R_{neg}, R_{cond}\right), \end{aligned} \tag{4}$$

where $E_{nr}$ denotes the negative rectifier. We employ the CFG reconstruction loss as the objective function to optimize the negative rectifier. The formulation is outlined below:

$$L_{cfg} = \|\boldsymbol{z_0} - \tilde{\boldsymbol{z}}_{\mathbf{0}}\|^2. \tag{5}$$

$\tilde{\boldsymbol{z}}_{\mathbf{0}}$ can be derived through the subsequent computational process:

$$\tilde{\boldsymbol{z}}_{\mathbf{0}} = \frac{1}{\sqrt{\bar{\alpha}_t}} z_t - \sqrt{\frac{1}{\bar{\alpha}_t} - 1} \cdot \tilde{\varepsilon}_\theta, \tag{6}$$

where

$$\begin{aligned} \tilde{\varepsilon}_\theta(z_t, t, R_{pos}, \hat{R}_{neg}, \hat{F}_{c,t}) = w \cdot \varepsilon_\theta(z_t, t, R_{pos}, \hat{F}_{c,t}) \\ + (1-w) \cdot \varepsilon_\theta(z_t, t, \hat{R}_{neg}, \hat{F}_{c,t}), \end{aligned} \tag{7}$$

and $R_{pos} = E_{text}(p)$ is the embedding of input prompt $p$.

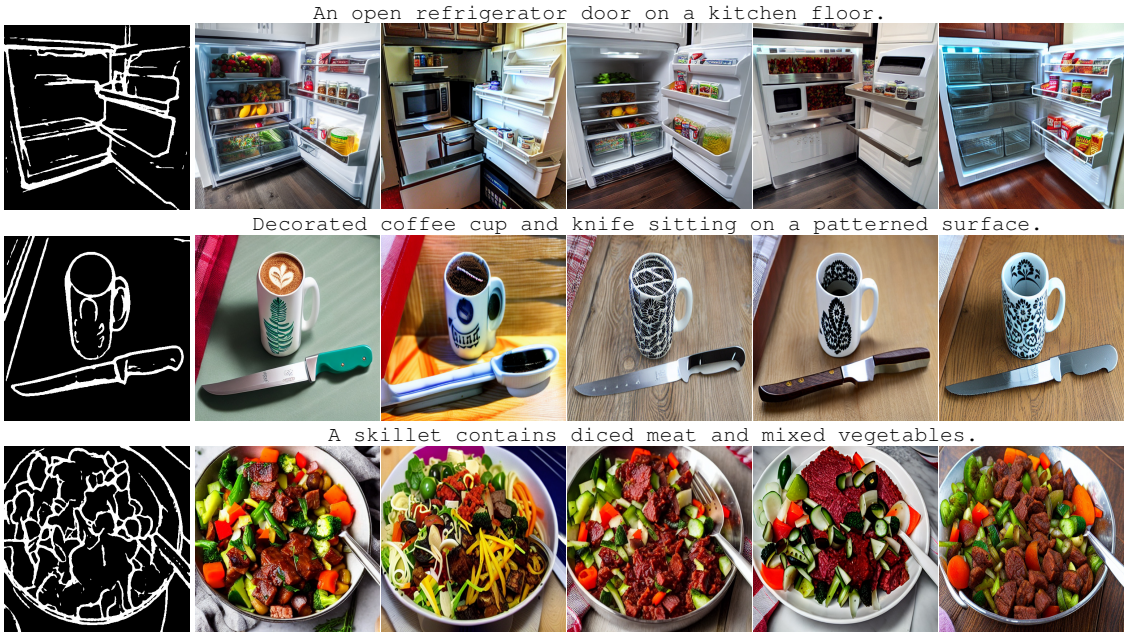
## 4. Experiments

In this section, we demonstrate that our method outperforms currently state-of-the-art controllable T2I models on few-shot conditional generation tasks (Sec. 4.2). The experiments are conducted with respect to five novel conditions (sketch, segmentation, pose, depth, and canny edge) on three datasets, which will be discussed in Sec. 4.1 together with evaluation metrics. Sec. 4.3 provides ablation studies on the proposed *prompt-free conditional learning* and *condition-specific negative rectification*.

### 4.1. Experiment settings

**Datasets and Metrics.** We evaluate our method on COCO [22], Human-Art [17], and instructPix2Pix [4] across five conditions. Conditional generation with sketch and segmentation are performed on the COCO dataset. COCO contains 118$K$ training images and 5$K$ testing images, which coupling with sketches and segmentation maps in its variants COCO17 [22] and COCO-stuff [5], respectively. For the pose-guided generation, we adopt Human-Art, which is an artistic dataset comprising 33.5$K$ training text-image-pose pairs and 4.5$K$ testing pairs from 19 scenes, both natural and artificial. For depth and canny edge, we use the data provided by instructPix2Pix, which has approximately 310$K$ image-text pairs. We manually separate the training and testing set of this dataset according to the label of each image, where the image that has a label starting with "0" is classified into the training set. The testing set is formed by 4.5$K$ randomly selected image-text pairs from the rest of the data. We employ Midas [32] to obtain the corresponding depth, while the corresponding Canny edge is acquired through the Canny edge detector [1]. To evaluate the performance, we use the conventional FID [11] to assess the generation quality. For measuring the degree of alignment between the generated content and provided conditions, we apply the SSIM [42] metric between the structure of the generated images with their ground truth conditions, denoted as cSSIM. In practice, we employ condition extraction networks to extract the structure for each generated image. We also use the AP metric [16] on Human-Art to measure the pose accuracy.

**Compared Methods,** including ControlNet-1.0 [44] (build upon SD-1.5) and T2IAdapter [26] (build upon SD-1.4) on all the five generation scenes. Additionally, we compare HumanSD [16] for the pose condition, and PromptDiff [43] for the conditions of segmentation, depth, and canny edge.




Conditions (a) ControlNet [44] (b) ControlNet-100 (c) T2IAdapter [26] (d) T2IAdapter-100 (e) Ours

Figure 5. Comparisons of sketch-guided text-to-image generations on COCO [22].


Conditions (a) ControlNet [44](b) ControlNet-100(c) T2IAdapter [26](d) T2IAdapter-100(e) PromptDiff [43](f) PromptDiff-100 (g) Ours

Figure 6. Comparisons of segmentation-guided text-to-image generations on COCO [22].

We retrained these models with only 100 training exemplar pairs under our setting, which are denoted with the suffix "-100". Besides, we also provide the original fully trained version of these models on the five evaluation conditions as a reference. For more details, please refer to the supplementary material.

**Implementation Details.** Following ControlNet, we build our model upon SD-1.5. For each evaluated condition, we train our model with 100 randomly sampled image-text-condition pairs from the training set of the three datasets, and evaluate on their whole testing set. During the training and testing phase, both the input images and conditions are resized to 512×512. The $\omega$ in Eq. 7 is set to 7.5. We adopt AdamW [24] as the optimizer in all our experiments. In the stage of prompt-free conditional learning, the learning rate is set to $5 \cdot 10^{-5}$ for adapting to segmentation and depth exemplars, while set as $1 \cdot 10^{-5}$ for other conditions. During the phase of condition-specific negative rectification, the learning rate is maintained at $1 \cdot 10^{-4}$ for all conditions. For each experiment, the training of our framework is finished within 4 hours on a single RTX 3090, with batch size 1.

| Condition / Dataset / Method | COCO [5, 22] | | | | InstructPix2Pix [4] | | | |
|---|---|---|---|---|---|---|---|---|
| | Sketch | | Segmentation | | Depth | | Canny edge | |
| | FID↓ | cSSIM↑ | FID↓ | cSSIM↑ | FID↓ | cSSIM↑ | FID↓ | cSSIM↑ |
| ControlNet [44] | 22.046 | 0.690 | 27.377 | 0.820 | 21.967 | 0.830 | 13.539 | 0.605 |
| T2IAdapter [26] | 19.445 | 0.683 | 24.254 | 0.811 | 15.856 | 0.792 | 16.167 | 0.441 |
| PromptDiff [43] | - | - | 35.837 | 0.815 | 20.202 | 0.823 | 28.084 | 0.504 |
| Control Net-100 | 27.598 | 0.709 | 31.109 | 0.828 | 28.396 | 0.764 | 34.596 | 0.432 |
| T2I-Adapter-100 | 21.053 | 0.65 | 22.103 | 0.821 | 20.379 | 0.769 | 19.096 | 0.429 |
| PromptDiff-100 | - | - | 27.194 | 0.816 | 23.896 | 0.721 | 20.148 | 0.472 |
| **Ours** | **21.049** | **0.692** | **20.726** | **0.835** | **19.137** | **0.803** | **16.710** | **0.475** |

Table 1. Quantitative comparisons on COCO [22] and InstructPix2Pix [4]. Top 2 records are marked in red and blue respectively.

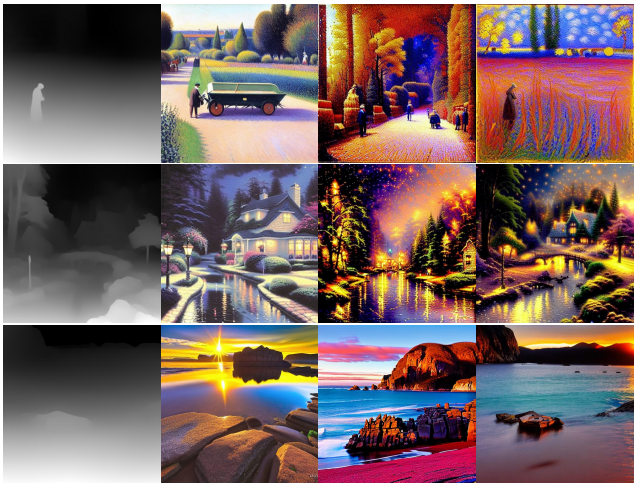



Figure 7. Comparisons of depth-guided generations on the InstructPix2Pix [4] dataset. The text prompts from the top row to the bottom row are "Image result for Henri Martin.", "Thomas Kinkade painter of light — Thomas Kinkade - Painter of Light - The Contrast Magazine.", and "Sunrise pirates bay, Tasmania by Robert-Todd.", respectively.

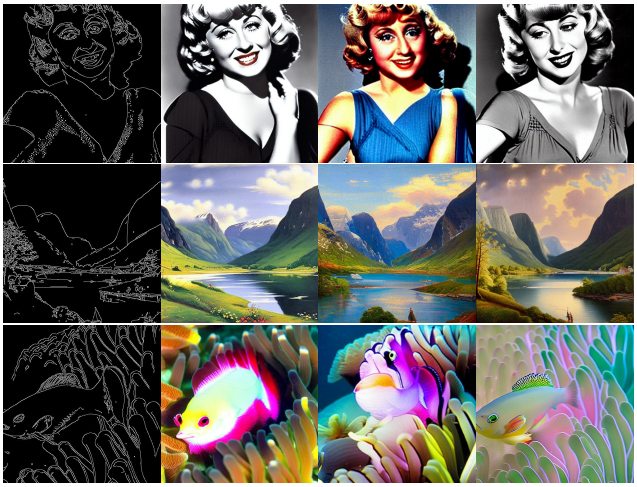



Figure 8. Comparisons of edge-guided generations on the InstructPix2Pix [4] dataset. The text prompts from the top row to the bottom row are "Joan Blondell (1906-1979), American actress, known for 'Grease.' Hollywood's Golden Age.", "A fjord in summer by Adelsteen Normann - reproduction oil painting.", and "Pink skunk anemone fish, Amphiprion perideraion, Fiji, natural history stock photograph.", respectively.

## 4.2. Comparisons

**Qualitative Experiments.** We visualize the generation results for each new condition and compare them with contrasting methods. Fig. 5 and Fig. 6 present visualization results conditioned on sketches and segmentation maps. Under a 100-sample setting (see Fig. 5 and Fig. 6), our approach exhibits strong visual coherence and consistency with the input conditions. The results of ControlNet-100 and T2IAdapter-100 show tight text-image relevance, but they fail to sustain structural consistency with the novel conditions as shown in Fig. 5b and Fig. 6f. Our approach, on the other hand, demonstrated significant consistency between the images and conditions. We have also provided extensive comparisons with with the full-trained methods, indicating our ability to produce high quality and authentic images. Visualizations of other conditions are presented in Fig. 7, Fig. 8 and Fig. 9. Kindly refer to the supplementary material for more results.

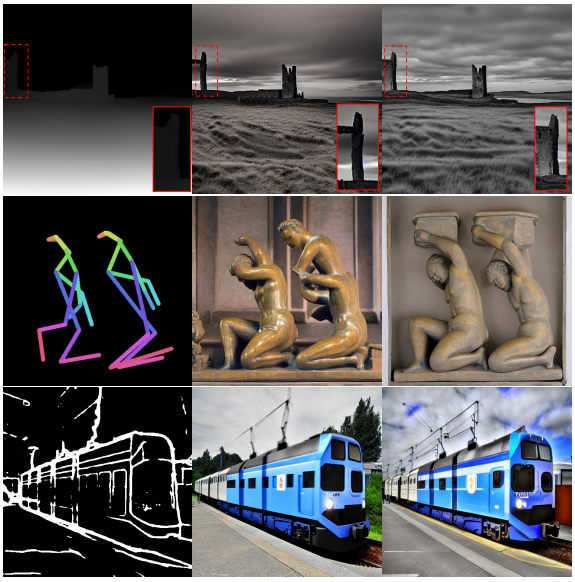
**Quantitative Experiments.** Quantitative results for different conditions, including sketch, segmentation maps, depth, Canny edge, and pose, are presented in Table 1 and Table 2. Under the 100-sample setting, our method exhibits a distinct advantage in both FID and cSSIM metrics, illustrating our superiority in terms of image quality and conditional control. Despite the few-shot comparison, we also evaluate the full version of comparative methods. As indicated in Table 1, our approach attains a significant level of performance that is on par with the full ControlNet, T2I-Adapter, and PromptDiff. In the case of segmentation, our few-sample approach achieves a 20.762 FID and 0.835 cSSIM, which substantially exceed the full-trained baselines.

## 4.3. Ablation Study

In this section, we perform ablation experiments for our proposed prompt-free conditional learning (PCL) and

Conditions ControlNet-100 HumanSD-100 Ours

Figure 9. Comparisons of pose-guided text-to-image image generations on the Human-Art [17] dataset. The text prompts from the top row to the bottom row are "Kids drawing, a painting of a rainbow and children walking in the grass.", "Watercolor, a painting of a man in uniform standing on a boat.", and "Relief, a stone carving of a man holding a cross.", respectively.



Conditions (a) PCL (b) PCL+CNR

Figure 10. Visualizations of ablation for our proposed PCL and CNR. The text prompts of the three generation task from the top to the bottom are "Dunstanburgh Mono by colin63.", "relief, two statues of men with their hands on their heads.", and "A blue and white train is moving on the rails.", respectively.

condition-specific negative rectification (CNR). We evaluate the outcomes of the two stages across three conditions: sketch, pose, and depth. Table 3 illustrates that the CNR could enhance the performance on both FID and cSSIM, indicating better image quality and conditional controllability. Visual enhancement can be observed clearly in Fig. 10. For the first case, CNR managed to amend the architectural misalignment. The second row demonstrates CNR's capacity to improve semantic alignment, enhancing the correspondence between generated semantics and the input pose. In the third example, CNR alleviates the excessive smoothing on the train carriage and the blur of overhead wires. Please refer to the supplementary for more ablation and analysis.

| Method | FID↓ | AP↑ | cSSIM↑ |
|---|---|---|---|
| ControlNet [44] | 40.768 | 36.43 | 0.852 |
| T2IAdapter [26] | 40.219 | 44.62 | 0.857 |
| HumanSD [16] | 36.817 | 47.51 | 0.863 |
| ControlNet-100 | 36.659 | 19.64 | 0.854 |
| T2IAdapter-100 | 42.601 | 13.80 | 0.851 |
| HumanSD-100 | 32.339 | 15.20 | 0.851 |
| Ours | **32.968** | **23.10** | **0.855** |

Table 2. Quantitative results for pose-guided generation on Human-Art [17]. Top 2 records are marked in red and blue respectively.

| Modules | | Sketch | | Pose | | | Depth | |
|---|---|---|---|---|---|---|---|---|
| PCL | CNR | FID↓ | cSSIM↑ | FID↓ | AP↑ | cSSIM↑ | FID↓ | cSSIM↑ |
| ✓ | | 21.245 | 0.684 | 36.334 | 23.10 | 0.853 | 19.299 | 0.803 |
| ✓ | ✓ | 21.049 | 0.692 | 32.968 | 23.10 | 0.855 | 19.137 | 0.803 |

Table 3. Ablation study on the proposed PCL and CNR.

## 5. Conclusion

We tackle the challenge of adapting the text-to-image generative model to novel diffusion conditions with limited training samples, particularly addressing their limitations in capturing detailed structural features when relying solely on text prompts. Our focus is on overcoming the issue of structural misalignment caused by imbalanced learning with sparse data. We introduce a two-stage optimization framework, comprising the prompt-free conditional learning and the condition-specific negative rectification, to reduce the text prompt bias and improve structural alignment. This approach significantly lowers the data requirements compared to existing methods, making it more feasible for real-world applications. Our framework demonstrates its effectiveness through extensive experimentation, proving its ability to adapt to new conditions efficiently with limited data.

**Acknowledgements.** The work is supported by China National Key R&D Program (No. 2023YFE0202700); Key-Area Research and Development Program of Guangzhou City (No.2023B01J0022); Guangdong International Technology Cooperation Project (No.2022A0505050009); National Natural Science Foundation of China (No.62302170); and Guangdong Natural Science Funds for Distinguished Young Scholar (No. 2023B1515020097); Singapore MOE Tier 1 Funds (MSS23C002); and the NRF Singapore under the AI Singapore Programme (No. AISG3-GV-2023-011).

## References


[1] Paul Bao, Lei Zhang, and Xiaolin Wu. Canny edge detection enhancement by scale multiplication. *IEEE TPAMI*, 27(9):1485–1490, 2005. 5

[2] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. 2023. 3

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 3

[4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, pages 18392–18402, 2023. 5, 7

[5] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *CVPR*, pages 1209–1218, 2018. 5, 7

[6] Guillaume Couairon, Marlène Careil, Matthieu Cord, Stéphane Lathuilière, and Jakob Verbeek. Zero-shot spatial layout conditioning for text-to-image diffusion models. In *ICCV*, pages 2174–2183, 2023. 3

[7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, volume 34, pages 8780–8794, 2021. 3

[8] Ziyi Dong, Pengxu Wei, and Liang Lin. Dreamartist: Towards controllable one-shot text-to-image generation via positive-negative prompt-tuning. *arXiv preprint arXiv:2211.11337*, 2022. 5

[9] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, pages 12873–12883, 2021. 2, 3

[10] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *CVPR*, pages 10696–10706, 2022. 3

[11] 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. In *NeurIPS*, volume 30, 2017. 5

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, volume 33, pages 6840–6851, 2020. 3

[13] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2, 3

[14] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778*, 2023. 3

[15] Yutao Jiang, Yang Zhou, Yuan Liang, Wenxi Liu, Jianbo Jiao, Yuhui Quan, and Shengfeng He. Diffuse3d: Wide-angle 3d photography via bilateral diffusion. In *CVPR*, pages 8998–9008, 2023. 3

[16] Xuan Ju, Ailing Zeng, Jianan Wang, Qiang Xu, and Lei Zhang. Human-art: A versatile human-centric dataset bridging natural and artificial scenes. In *ICCV*, 2023. 2, 3, 5, 8

[17] Xuan Ju, Ailing Zeng, Jianan Wang, Qiang Xu, and Lei Zhang. Human-art: A versatile human-centric dataset bridging natural and artificial scenes. In *CVPR*, pages 618–629, 2023. 5, 8

[18] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *CVPR*, pages 10124–10134, 2023. 2

[19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 2

[20] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186, 2019. 3

[21] Sungnyun Kim, Junsoo Lee, Kibeom Hong, Daesik Kim, and Namhyuk Ahn. Diffblender: Scalable and composable multimodal text-to-image diffusion models. *arXiv preprint arXiv:2305.15194*, 2023. 3

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 5, 6, 7

[23] Haofeng Liu, Chenshu Xu, Yang Yifei, Zeng LiHua, and He Shengfeng. Drag your noise: Interactive point-based editing via diffusion semantic propagation. In *CVPR*, 2024. 3

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2018. 6

[25] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *CVPR*, pages 6038–6047, 2023. 5

[26] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 1, 2, 3, 4, 5, 6, 7, 8

[27] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, pages 8162–8171. PMLR, 2021. 3

[28] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, pages 16784–16804. PMLR, 2022. 3

[29] Xingang Pan, Ayush Tewari, Thomas Leimkühler, Lingjie Liu, Abhimitra Meka, and Christian Theobalt. Drag your gan: Interactive point-based manipulation on the generative image manifold. In *ACM SIGGRAPH*, pages 1–11, 2023. 3

[30] 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. In *NeurIPS*, 2023. 3

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

[32] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE TPAMI*, 44(3):1623–1637, 2020. 5

[33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2, 3

[34] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015. 3

[35] 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. In *NeurIPS*, volume 35, pages 36479–36494, 2022. 3

[36] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, volume 35, pages 25278–25294, 2022. 3

[37] Yujun Shi, Chuhui Xue, Jiachun Pan, Wenqing Zhang, Vincent YF Tan, and Song Bai. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. *arXiv preprint arXiv:2306.14435*, 2023. 3

[38] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, pages 2256–2265. PMLR, 2015. 3

[39] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2020. 3

[40] Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. Sketch-guided text-to-image diffusion models. In *ACM SIGGRAPH*, pages 1–11, 2023. 3

[41] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation. *arXiv preprint arXiv:2205.12952*, 2022. 3

[42] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004. 5

[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. 3, 5, 6, 7

[44] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 1, 2, 3, 5, 6, 7, 8

[45] Zongyan Zhang, Haohan Weng, Tong Zhang, and C. L. Philip Chen. A broad generative network for two-stage image outpainting. *IEEE TNNLS*, pages 1–15, 2023. 3

[46] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. In *NeurIPS*, 2023. 3

[47] Chenxi Zheng, Bangzhen Liu, Xuemiao Xu, Huaidong Zhang, and Shengfeng He. Learning an interpretable stylized subspace for 3d-aware animatable artforms. *IEEE TVCG*, 2024. 3

[48] Chenxi Zheng, Bangzhen Liu, Huaidong Zhang, Xuemiao Xu, and Shengfeng He. Where is my spot? few-shot image generation via latent subspace optimization. In *CVPR*, pages 3272–3281, 2023. 3