# Simplifying Control Mechanism in Text-to-Image Diffusion Models

**Zhida Feng**[1,2,3], **Li Chen**[1,2,*], **Yuenan Sun**[1,2], **Jiaxiang Liu**[3], **Shikun Feng**[3]

[1]School of Computer Science and Technology, Wuhan University of Science and Technology
[2]Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan University of Science and Technology
[3]Baidu Inc.

## Abstract

ControlNet has significantly advanced controllable image generation by integrating dense conditions (such as depth and canny edges) with text-to-image diffusion models. However, ControlNet's integration requires an additional amount nearly equal to half of the base diffusion model's parameters, making it inefficient. To address this, we introduce Simple-ControlNet, an efficient and streamlined network for controllable text-to-image generation. It employs a single-scale projection layer to incorporate condition information into the denoising U-Net. It is supplemented by Low-Rank Adapter (LoRA) parameters to facilitate condition learning. Impressively, Simple-ControlNet requires fewer than 3 million parameters for the control mechanism, substantially less than the 300 million needed by ControlNet. Our extensive experiments confirm that Simple-ControlNet matches and surpasses ControlNet's performance across a broad range of tasks and base diffusion models, showcasing its utility and efficiency.

**Code** — https://github.com/feng-zhida/Simple-ControlNet

## Introduction

The field of image generation has made significant strides with deep generative models, particularly diffusion models (Sohl-Dickstein et al. 2015; Song et al. 2020; Ho, Jain, and Abbeel 2020). These models have unlocked new possibilities to generate highly realistic and diverse images, pushing the frontiers of visual synthesis. However, the quest for controlled image generation, which allows precise manipulation of generated content to meet specific user requirements, remains a challenge.

Recent advances in text-to-image diffusion models (Chen et al. 2023; Ramesh et al. 2022; Feng et al. 2023; Saharia et al. 2022b; Rombach et al. 2022; Podell et al. 2023; Pernias, Rampas, and Aubreville 2023) have markedly improved the controllability of image generation, facilitating the creation of images closely aligned with user-provided text prompts. Despite their success, relying solely on textual descriptions often fails to convey the detailed controls required for precise image generation. ControlNet (Zhang, Rao, and Agrawala 2023) emerged as a revolutionary approach by integrating detailed conditions, such as segmentation maps and edge maps, with text-to-image models to improve controllability. However, ControlNet introduced substantial complexity, necessitating nearly half the parameters of the base U-Net model, complicating training and deployment.

Addressing these challenges, we introduce Simple-ControlNet, a streamlined and efficient architecture for controllable text-to-image generation. Unlike previous control mechanisms (Zhang, Rao, and Agrawala 2023; Mou et al. 2024) that rely on condition encoders injecting information at multiple scales, Simple-ControlNet employs a simpler, single-scale approach. Specifically, we use a lightweight projection block consisting of 8 convolutional layers to integrate condition information directly into the U-Net's top-level features. This design substantially reduces complexity while still allowing the model to effectively leverage condition inputs. At the same time, Low-Rank Adapter (LoRA) parameters facilitate learning condition information embedded in the hidden states. This approach significantly simplifies the integration of control conditions into the diffusion model.

This paper makes the following contributions.

- We introduce Simple-ControlNet, a model that drastically reduces the parameter count needed for control mechanisms by nearly a hundredfold (from 344.5M to 2.7M), simplifying deployment and reducing training complexity.
- Through extensive experiments, we demonstrate that Simple-ControlNet matches and surpasses ControlNet's performance across multiple dimensions, including qualitative and quantitative comparisons, efficiency analyses, and human evaluations.
- Demonstration of Simple-ControlNet's versatility across a diverse array of tasks such as depth-to-image, boundary-to-image conversion, and advanced image processing techniques including inpainting, outpainting, and Super-Resolution.

## Related Work

**Image-to-Image Translation** Numerous GAN-based image-to-image translation methods (Choi et al. 2018,

*Corresponding Author

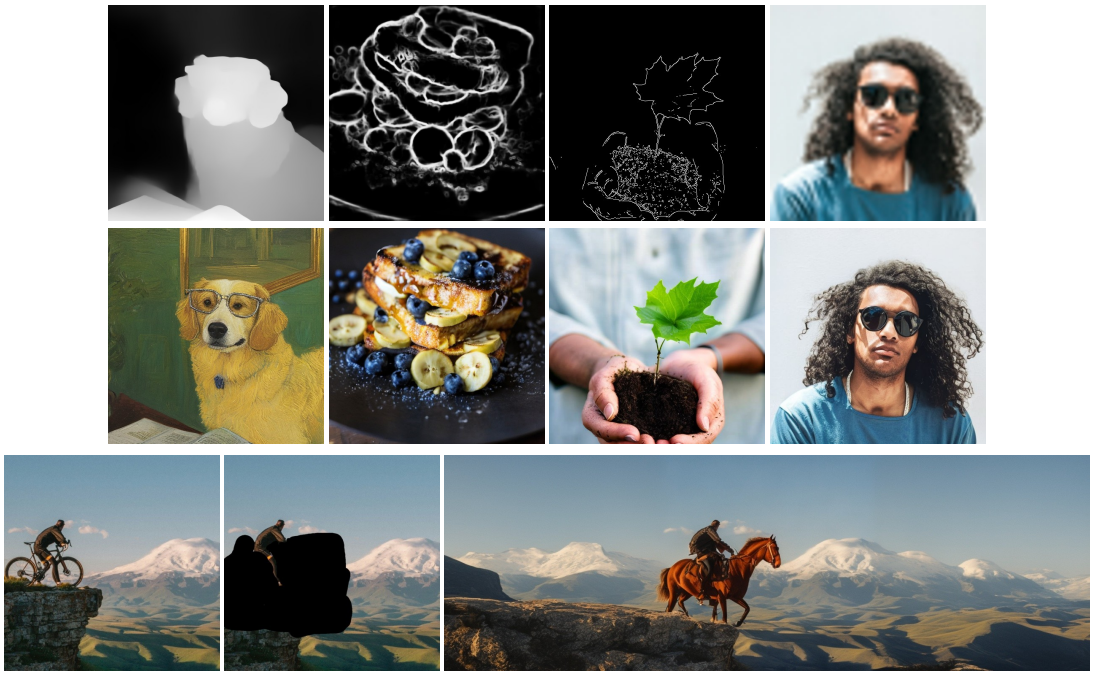
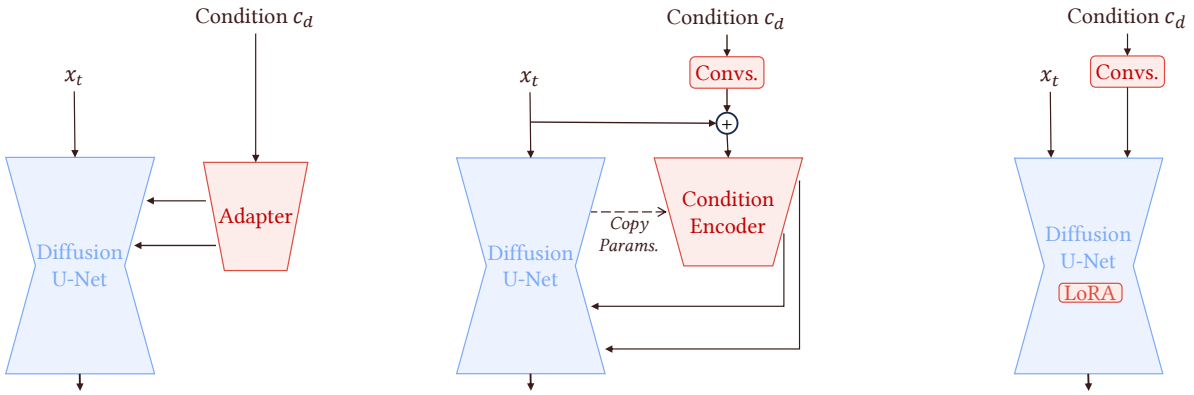


Figure 1: Selected Simple-ControlNet samples with various conditions, showcasing tasks including Depth-to-Image, Boundary-to-Image, Edge-to-Image, and Super-Resolution. The inpainting and outpainting tasks are jointly performed within a single pipeline, using the prompt "A person riding a horse on a cliff."

2020; Isola et al. 2017; Wang et al. 2018; Park et al. 2019; Zhu et al. 2017; Taigman, Polyak, and Wolf 2017; Richardson et al. 2021) have been extensively studied in this field. Those approaches typically learn the mapping from the source to the target image domain. Methods can be categorized as supervised or unsupervised (Zhu et al. 2017), paired or unpaired (Taigman, Polyak, and Wolf 2017). The autoregressive transformer methods (Ramesh et al. 2021; Esser, Rombach, and Ommer 2021) consider the discrete tokens of the source domain image as the initial part of the input sequence and then decode the discrete tokens of the target domain image. Diffusion-based approaches (Wang et al. 2022; Saharia et al. 2022a; Ramesh et al. 2022; Saharia et al. 2022b) commonly incorporate the image of the source domain as an additional condition for the denoising network during diffusion training. Recently, several methods (Zhang, Rao, and Agrawala 2023; Mou et al. 2024; Zhao et al. 2024) have explored novel architectures that enhance the image-to-image translation capabilities of pre-trained diffusion models while preserving their inherent generative power.

**Text-to-Image Diffusion Models** Diffusion models (Song et al. 2020; Sohl-Dickstein et al. 2015; Ho, Jain, and Abbeel 2020) have recently emerged as a significant advancement in image generation, impressing their ability to produce high-quality images. Using large-scale text-to-image datasets for training, text-to-image diffusion models (Nichol et al. 2021; Ramesh et al. 2022; Rombach et al. 2022; Feng et al. 2023; Saharia et al. 2022b; Chen et al. 2023) demonstrate an extraordinary ability to generate high-fidelity images and customize these images based on input text. Many models (Nichol et al. 2021; Ramesh et al. 2022; Saharia et al. 2022b) perform diffusion in the pixel space and employ cascading diffusion techniques (Ho et al. 2022) to produce high-resolution images. On the other hand, Latent Diffusion Models (LDM) adopt a distinct approach by training neural networks to reduce the dimensionality of images. By conducting diffusion in latent space, LDMs significantly reduce computational demands while still allowing the generation of high-resolution images through a single neural network. Following the success of LDM, many subsequent text-to-image diffusion models (Feng et al. 2023; Chen et al. 2023) have adopted the approach of performing diffusion in latent space.

**Controlling Text-to-Image Diffusion Models** Recently, some approaches have attempted to incorporate dense conditional information into frozen pre-trained text-to-image diffusion models. The T2I-adapter (Mou et al. 2024) trains an adapter that outputs multiscale feature maps and adds them to the U-Net of Stable Diffusion. GLIGEN (Li et al. 2023) designs cross-attention layers and inserts them into the network to incorporate conditional information. ControlNet (Zhang, Rao, and Agrawala 2023) duplicates the parameters of the Denoising U-Net encoder to create a Condition Encoder, whose outputs containing control information are fed as multiscale features into the Denoising U-Net decoder. Uni-ControlNet integrates many control conditions into a single control network. In contrast to these




(a) T2I-Adapter (b) ControlNet (c) Simple-ControlNet (Ours)

Figure 2: Comparison of three control mechanisms. (a) T2I-Adapter trains an Adapter from scratch, which takes the condition as input and outputs multiscale features to the U-Net's encoder; (b) ControlNet duplicates the parameters of the U-Net Encoder, taking $x_t$ and the condition as inputs, and outputs multiscale features to the U-Net's decoder; (c) Simple-ControlNet (Ours) directly inputs the condition to the U-Net through a simple projection layer, and then introduces additional LoRA parameters to learn the extra condition information in the hidden state. The red and blue modules represent whether parameters are updated or not updated during training, respectively.

methods, we propose a simple approach to inject control information by directly inserting conditions into the top layer of the U-Net through a few convolutional layers. We then employ Low-Rank Adaptation (LoRA) (Hu et al. 2021) on the self-attention layers to handle additional conditional hidden states. Our method offers a straightforward and effective way to incorporate control information into text-to-image diffusion models.

**Neural Networks Fine-tuning** Fine-tuning large pre-trained models on downstream tasks has become increasingly common (Ruiz et al. 2023; Devlin et al. 2019; Ouyang et al. 2022). As a fine-tuning technique, the adaptation method (Hu et al. 2021; Pfeiffer et al. 2021; Houlsby et al. 2019) has received a lot of attention in natural language processing. Techniques such as LoRA (Hu et al. 2021) and orthogonal fine-tuning (Qiu et al. 2023) have demonstrated the potential to efficiently incorporate new capabilities into existing models without compromising their original strengths. This approach has been particularly influential in image generation, where pre-trained models like Stable Diffusion (Rombach et al. 2022) are fine-tuned with additional inputs or constraints to achieve specific generative outcomes. This strategy aligns with our exploration of more effective and efficient methods for integrating dense conditions into the image generation process, aiming to enhance control while minimizing the additional computational burden.

## Method

### Background

Denoising Diffusion Probabilistic Models (DDPMs) (Ho, Jain, and Abbeel 2020) are score-based generative models that have recently been used in image generation. These models utilize a diffusion process that incrementally introduces diagonal Gaussian noise into an initial data sample, $x$, converting it into an isotropic Gaussian distribution over $T$ steps as follows:

$$x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1-\alpha_t}\epsilon_t, , t \in \{1, \dots, T\}, \quad (1)$$

where $x_0 = x$, $x_T \sim \mathcal{N}(0, I)$, $\epsilon_t \sim \mathcal{N}(0, I)$, and $\{\alpha_t\}_{t=1}^T$ is a predefined schedule.

The forward process allows for the sampling of $x_t$ at any timestep $t$ in closed form. Defining $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, we obtain:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t} x_0, (1-\bar{\alpha}_t)I), \quad (2)$$

This can also be expressed as:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, , \text{where } \epsilon \sim \mathcal{N}(0, I), \quad (3)$$

Subsequently, a neural network denoted as $\theta$ is used to predict $\epsilon$ (potentially outputting $x_0$ or $v$, then reparameterized to $\epsilon$) to improve the controllability of generation. Caption information and dense conditions are integrated into the inputs of the neural network, enabling the model to predict $x_0$ at step $t$ as follows:

$$\hat{x}_{0,t} = \frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t, t, c_t, c_d)), \quad (4)$$

where $c_t$ and $c_d$ represent text and dense condition information, respectively.

In the inference phase of DDPMs, since $x_0$ is unknown, the model iteratively generates $x_{t-1}$ based on $x_t$ and $\hat{x}_{0,t}$:

$$\begin{aligned} x_{t-1} =& \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\sqrt{\alpha_t} x_t \\ &+ \frac{1-\alpha_t}{1-\bar{\alpha}_t}\sqrt{\bar{\alpha}_{t-1}}\hat{x}_{0,t} \\ &+ \sqrt{\frac{(1-\bar{\alpha}_{t-1})(1-\alpha_t)}{1-\bar{\alpha}_t}}\epsilon'_t, \end{aligned} \quad (5)$$

where $\epsilon'_t \sim \mathcal{N}(0, I)$ is sampled Gaussian noise and $t \in \{T, \dots, 1\}$.

## Designing Simple-ControlNet

Recent methods for controlling pre-trained text-to-image diffusion models, such as T2I-Adaptar (Mou et al. 2024) and ControlNet (Zhang, Rao, and Agrawala 2023), utilize an additional condition encoder to insert multiscale features into the frozen U-Net's intermediate hidden states (refer to Figure 2a and Figure 2b). These condition encoders match the depth and width of the Diffusion U-Net encoder or decoder, adding significant complexity to the model.

In this section, we introduce Simple-ControlNet Figure 2c. This streamlined approach employs a single-scale projection layer at the top of the U-Net to embed conditions, simplifying the architecture. Using the LoRA technique enables us to learn this conditional information, effectively reducing network complexity.

**Single-Scale Projection** We use a shallow projection layer (with only 8 convolutional layers to map the condition to a feature map with the same shape as the U-Net's top-level features.

Let $F(\cdot;\Phi)$ be our projection layer and $h_0 \in \mathbb{R}^{H\times W\times C}$ be the top-level features of the U-Net.

F can map the dense condition $c_d$ to a feature map $y$ through parameters $\Phi$:

$$y = F(c_d;\Phi),, y \in \mathbb{R}^{H\times W\times C}, \tag{6}$$

We then add $y$ to $h_0$ to introduce the condition information into the U-Net:

$$\hat{h_0} = h_0 + \lambda_P \cdot y, \tag{7}$$

where $h_0$ is the zeroth layer of the network (i.e., the top-level features), and $\lambda_P$ is a scalable factor set to 1 during training. Since only single-scale features are required, the network depth can be independent of the U-Net, and the top-level features have the least width, further reducing the number of parameters.

**Learning Condition Information** Our projection layer scale is small, so it cannot output features with rich semantic information. Therefore, we need additional parameters to understand the condition information embedded in the hidden states. Simple-ControlNet addresses this using LoRA (Low-Rank Adaptation) (Hu et al. 2021). Consider a layer updated by LoRA with weight $W \in \mathbb{R}^{d\times d}$. We use a low-rank matrix:

$$\Delta W = \frac{\alpha}{r} BA,, \text{where } B \in \mathbb{R}^{d\times r}, A \in \mathbb{R}^{r\times d}, \tag{8}$$

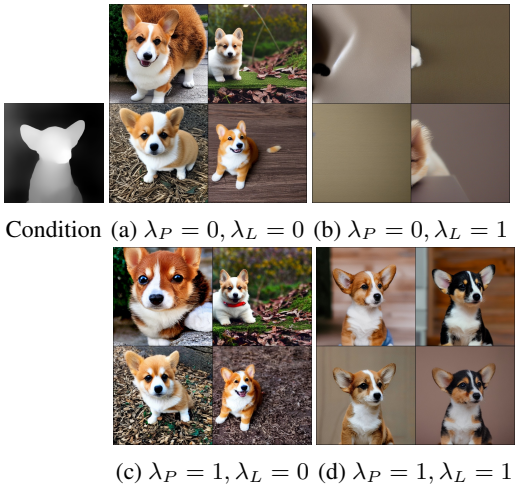
where $r$ and $\alpha$ are pre-defined hyperparameters. Given an input hidden state $h$, we have:

$$z = h(W + \lambda_L \cdot \Delta W) + b, \tag{9}$$

This can be decomposed into:

$$z = h \cdot W + \lambda_L \cdot h \cdot \Delta W + b, \tag{10}$$

where $z$ is the output of this linear layer, and $\lambda_L$ is a scalable factor set to 1 during training. The condition information is embedded in $h$. Since $W$ is not updated during training, we hypothesize that the condition information can be learned by updating the parameters $\Delta W$.



Condition (a) $\lambda_P = 0, \lambda_L = 0$ (b) $\lambda_P = 0, \lambda_L = 1$ (c) $\lambda_P = 1, \lambda_L = 0$ (d) $\lambda_P = 1, \lambda_L = 1$

Figure 3: Results with different $\lambda$ values using same sampled noise. The $\lambda_P$ and $\lambda_L$ are used to control the state of the projection layer and LoRA, respectively, where 0 indicates deactivated and 1 indicates activated.

By controlling the activation of the projection layer and the LoRA layer in a pre-trained Simple-ControlNet, we obtained Figure 3. We observed that when only the projection layer is activated (Figure 3a vs. Figure 3c), the variation in the results generated is minimal. In contrast, activating the LoRA layer enables the model to generate images relevant to the condition (Figure 3c vs. Figure 3d). Furthermore, the generation of non-meaningful images in Figure 3b highlights that LoRA parameters are tightly coupled with condition information during training. This suggests that the projection layer primarily embeds the condition within our architecture, while the LoRA layer interprets and utilizes this condition information.

## Training Simple-ControlNet

**Parameter Initialization** The initialization of parameters plays a crucial role in the successful integration of new conditional layers, particularly when fine-tuning diffusion models. In accordance with ControlNet's findings (Zhang, Rao, and Agrawala 2023), we set the weights and biases of the final layer in the projection layers to zero. This strategy minimizes the initial impact of control information on hidden states, promoting a smoother adaptation process. Similarly, we zero-initialize the matrix $A$ in the Lora module, which is equivalent to zero-initializing $\Delta W$.

**Training Objective** For $\epsilon$ prediction models, such as Stable Diffusion v1-5, the training loss is defined as:

$$\mathcal{L} = ||\epsilon - \theta(x_t, t, c, d)||^2, \tag{11}$$

For models that utilize v prediction, like Stable Diffusion v2-1, the loss is calculated by:

$$\mathcal{L} = ||\tilde{\mu}_t - \theta(x_t, t, c, d)||^2, \tilde{\mu}_t = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon). \tag{12}$$

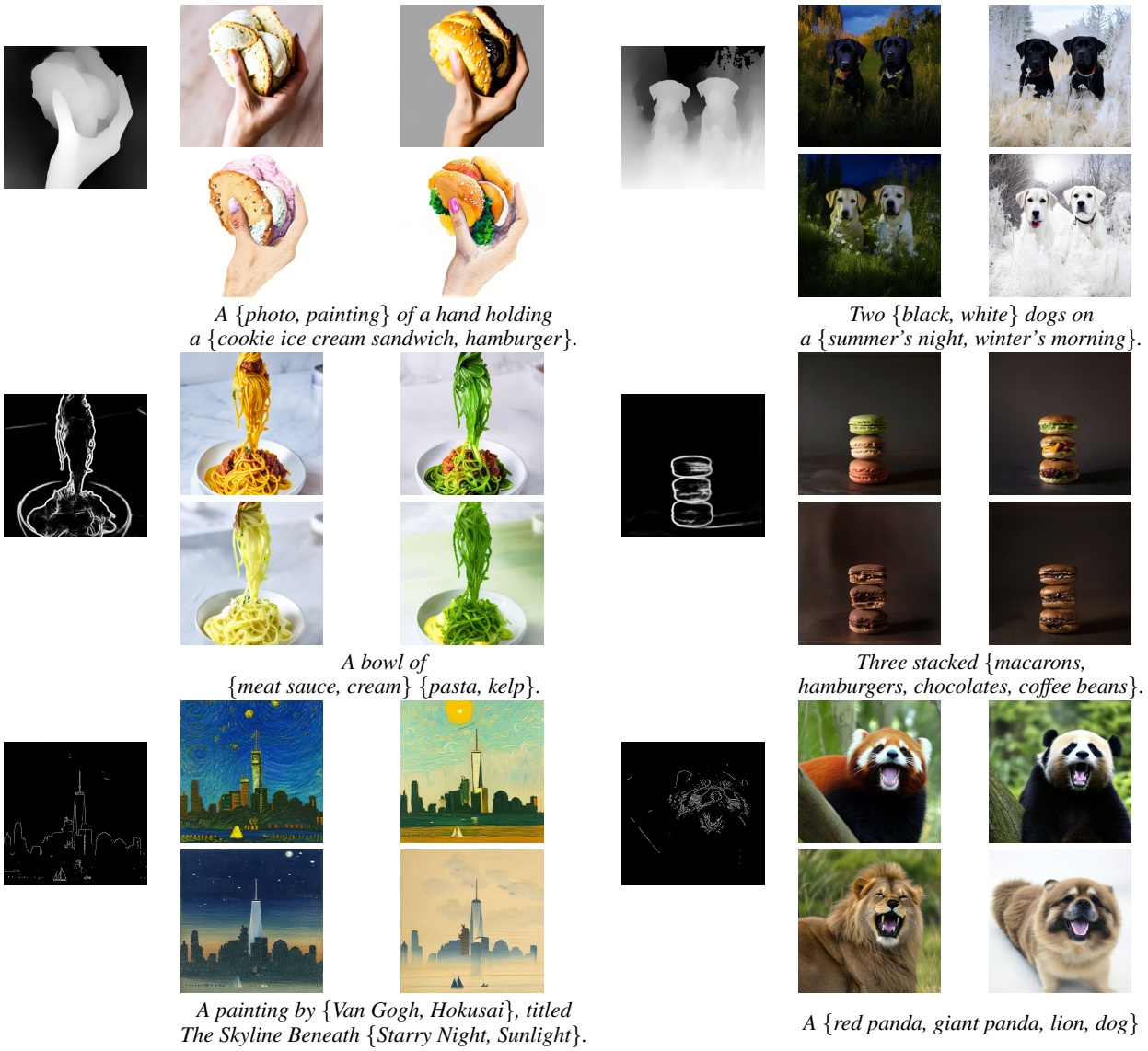



Figure 4: Results of Simple-ControlNet under various control conditions and prompts.

## Experiments

**Implementation Details** We sampled 2 million text-image pairs from the COYO-700M (Byeon et al. 2022) dataset for training. All models have trained over 40,000 iterations with a batch size of 128 using the AdamW (Loshchilov and Hutter 2019) optimizer, with settings $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We applied LoRA (Low-Rank Adaptation) (Hu et al. 2021) across all self-attention layers, employing a rank of 8, and set the LoRA dropout to 0.1. This adaptation, along with single-scale projection layers, added additional 1.6M and 1.1M parameters, respectively, to the pre-trained Stable Diffusion model (Rombach et al. 2022).

**Qualitative Evaluation** We use Stable Diffusion v1-5 (Rombach et al. 2022) as the base model, assessing our method with various controlled inputs such as Depth (Ranftl et al. 2022), HED (Xie and Tu 2015) Boundary and Canny (Canny 1986) Edge. All models, including ours, use the DPM-Solver (Lu et al. 2022) configured for 25 steps with a control strength of 1.0. Figure 4 demonstrates the results of Simple-ControlNet under various control conditions and prompts. Simple-ControlNet performs well on diverse prompts, effectively performing tasks that include entity transformation and style transfer. Figure 5 showcases the qualitative comparison results, which align with the quantitative findings. Simple-ControlNet exhibits superior adherence to the condition, as exemplified by the hippo case in the first row, where Simple-ControlNet accurately generates an image of a hippo with its mouth open, while ControlNet fails to do so. Furthermore, ControlNet suffers from apparent overexposure issues under the HED boundary condition (second row), whereas Simple-ControlNet produces more photorealistic images. Lastly, Simple-ControlNet excels in generating human faces, as observed in the last row.

| | Model | Extra Params. | Depth | | HED Boundary | | | Canny Edge | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | FID ↓ | CLIP Score ↑ | RMS↓ | FID↓ | CLIP Score ↑ | RMS↓ | FID ↓ | CLIP Score ↑ | SSIM ↑ |
| w/o CFG | T2I-Adapter (Mou et al. 2024) | 73.4M | 27.57 | 0.2857 | 0.1593 | - | - | - | 29.40 | 0.2810 | 0.4528 |
| | ControlNet (Zhang, Rao, and Agrawala 2023) | 344.5M | 21.28 | 0.2895 | 0.1372 | 41.75 | 0.2823 | 0.1942 | 26.58 | 0.2879 | 0.4333 |
| | ControlNet v1.1 (Zhang, Rao, and Agrawala 2023) | 344.5M | 21.13 | 0.2876 | 0.1378 | - | - | - | 17.82 | 0.2942 | 0.5534 |
| | Uni-ControlNet (Zhao et al. 2024) | 437.8M | 46.32 | 0.2625 | 0.2649 | 20.26 | 0.2917 | 0.1633 | 25.99 | 0.2880 | 0.4759 |
| | Simple-ControlNet (Ours) | **2.7M** | **15.59** | **0.2949** | **0.1326** | **10.11** | **0.3025** | **0.1384** | **12.93** | **0.3065** | **0.5758** |
| w/ CFG = 7.5 | T2I-Adapter (Mou et al. 2024) | 73.4M | 12.09 | **0.3159** | 0.1580 | - | - | - | 10.49 | 0.3127 | 0.4676 |
| | ControlNet (Zhang, Rao, and Agrawala 2023) | 344.5M | 11.86 | 0.3148 | 0.1342 | 11.92 | **0.3147** | 0.1347 | 10.63 | 0.3140 | 0.4631 |
| | ControlNet v1.1 (Zhang, Rao, and Agrawala 2023) | 344.5M | 12.61 | 0.3147 | 0.1349 | - | - | - | 8.93 | **0.3156** | 0.5391 |
| | Uni-ControlNet (Zhao et al. 2024) | 437.8M | 12.88 | 0.3129 | 0.2409 | 10.63 | 0.3120 | 0.1654 | 9.96 | 0.3143 | 0.4894 |
| | Simple-ControlNet (Ours) | **2.7M** | **10.10** | 0.3147 | **0.1317** | **7.97** | 0.3133 | **0.1344** | **8.71** | 0.3139 | **0.5740** |

Table 1: Quantitative comparison with other models. We present results both without and with Classifier-Free Guidance (CFG). When using CFG, we set the guidance scale to 7.5 for all models, which is a default setting in Stable Diffusion.


Figure 5: Qualitative comparison with ControlNet.

**Quantitative Evaluation** We conducted a quantitative comparison for three tasks using T2I-Adapter (Mou et al. 2024), ControlNet (Zhang, Rao, and Agrawala 2023), and Uni-ControlNet (Zhao et al. 2024). Our evaluation set comprised 10,000 image-text pairs sampled from the COCO (Lin et al. 2014) val2014 dataset. We used several metrics for assessment: the Fréchet Inception Distance (FID) (Heusel et al. 2017) to evaluate image quality; the CLIP (Radford et al. 2021)-Score using ViT-B/32 to assess image-text alignment; and both RMS and SSIM for evaluating the consistency of images with their conditions. The results, as shown in the Table 1, indicate that Simple-ControlNet excelled in image quality, image-text alignment, and image-condition consistency without classifier-free guidance (Ho and Salimans 2022). However, when the classifier-free guidance was set to 7.5, it also achieved the best performance in terms of quality and condition consistency, showing comparable results in CLIP-Score.

**Human Preference Study** We collected a set of 300 images from the Internet. For each image, we generated captions using InstructBlip (Dai et al. 2023), resulting in 300 image-text pairs. These pairs serve as the validation set for human evaluation. For each model and each condition, four images will be generated for each prompt. Five participants assessed the models in three dimensions: Image Fidelity, Text-Image Alignment, and Image-Condition Alignment. They chose the better image, or declared a tie, from a mixed sequence of outcomes from both our model and a competitor. This selection process aimed to calculate a preference rate as a metric. The findings, illustrated in Figure 6, reveal Simple-ControlNet's significant lead in Image Fidelity and slight advantages in alignment aspects, showing comparable alignment levels across most samples when compared with other models, succinctly emphasizing our model's effectiveness in producing and aligning images with their textual descriptions and specified conditions.

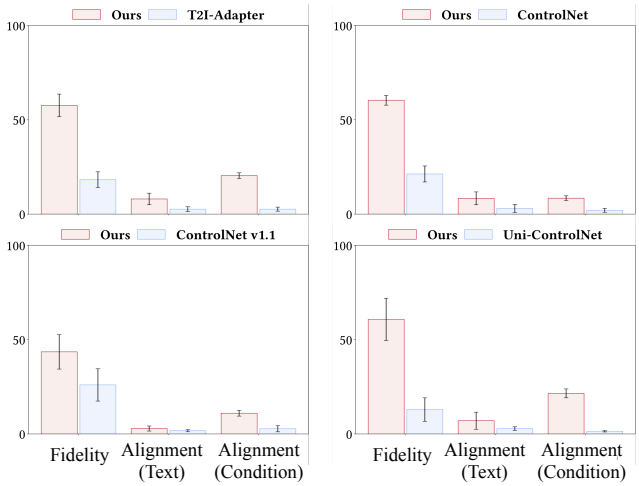



Figure 6: Human Preference Results.

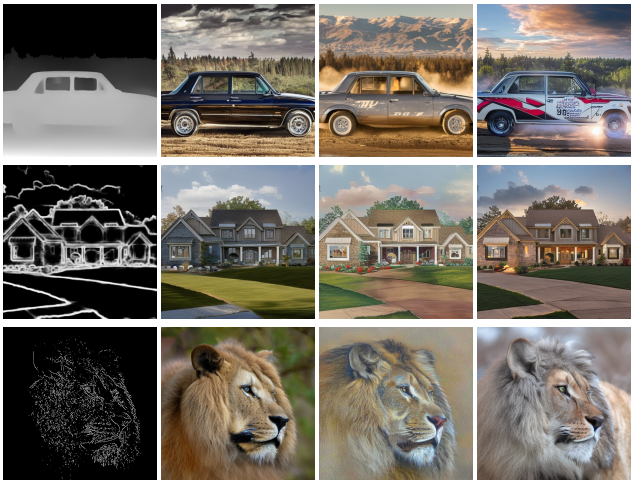


Figure 7: Non-Prompt Test (NPT) results showcasing Simple-ControlNet's ability to understand and generate images based solely on conditionings without prompts.

**Non-Prompt Test** We observed that ControlNet has released an additional ablation study on its GitHub discussions page [1]. The Non-Prompt Test (NPT) was introduced to evaluate whether a model can generate semantically meaningful images solely from the condition input, without relying on any text prompt. ControlNet relies on a large condition encoder to achieve this, but Figure 7 demonstrates that Simple-ControlNet also passes NPT while using far fewer additional parameters. This result highlights Simple-ControlNet's efficiency and its robust capability to understand dense conditions independently.

**User Input** We conducted a simple test to assess the adaptability of the Simple-ControlNet model to user input. The user provided an edge image with a width of one pixel, and we used the Canny edge version of the model to generate the corresponding image. The results are shown in Figure 8, demonstrating that the Simple-ControlNet model effectively adapts to user inputs.

[1] https://github.com/lllyasviel/ControlNet/discussions/188

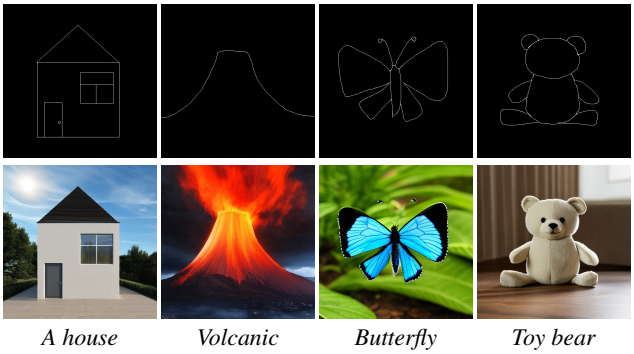



Figure 8: Results upon inputting a user-drawn edge.


Figure 9: Transfer pre-trained Simple-ControlNet to community models without training the neural networks again. The prompts for the first and second lines are "house" and "an old man" respectively.

**Transfer to other community models.** Part of ControlNet's usability stems from its transferability; it requires only a single pre-training on a base model (e.g., Stable Diffusion 1.5) and can then be directly applied to other fine-tuned models without further optimization. To demonstrate that Simple-ControlNet possesses the same characteristic, we directly applied our Simple-ControlNet (pre-trained on SD1.5) to four community models ( Figure 9) without any additional fine-tuning. The results indicate that Simple-ControlNet similarly exhibits this straightforward plug-and-play transferability, further validating its usability.

## Conclusion

In this work, we have introduced Simple-ControlNet, a streamlined and efficient architecture for controllable text-to-image generation. By simplifying the insertion of condition information, transitioning from a multiscale to a single-scale layer, and employing LoRA to learn the control information embedded in the hidden states, Simple-ControlNet demonstrates not only a significant reduction in the additional parameter count, but also superior performance in generating images that closely align with both textual and dense condition inputs.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (62271359).

## References

Byeon, M.; Park, B.; Kim, H.; Lee, S.; Baek, W.; and Kim, S. 2022. COYO-700M: Image-Text Pair Dataset. https://github.com/kakaobrain/coyo-dataset.

Canny, J. F. 1986. A Computational Approach to Edge Detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8(6): 679–698.

Chen, J.; Yu, J.; Ge, C.; Yao, L.; Xie, E.; Wu, Y.; Wang, Z.; Kwok, J. T.; Luo, P.; Lu, H.; and Li, Z. 2023. PixArt-$\alpha$: Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis. *CoRR*, abs/2310.00426.

Choi, Y.; Choi, M.; Kim, M.; Ha, J.-W.; Kim, S.; and Choo, J. 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8789–8797.

Choi, Y.; Uh, Y.; Yoo, J.; and Ha, J.-W. 2020. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8188–8197.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. C. H. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 4171–4186.

Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12873–12883.

Feng, Z.; Zhang, Z.; Yu, X.; Fang, Y.; Li, L.; Chen, X.; Lu, Y.; Liu, J.; Yin, W.; Feng, S.; et al. 2023. Ernie-vilg 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10135–10145.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 6626–6637.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. volume 33, 6840–6851.

Ho, J.; Saharia, C.; Chan, W.; Fleet, D. J.; Norouzi, M.; and Salimans, T. 2022. Cascaded Diffusion Models for High Fidelity Image Generation. *J. Mach. Learn. Res.*, 23: 47:1–47:33.

Ho, J.; and Salimans, T. 2022. Classifier-Free Diffusion Guidance. *CoRR*, abs/2207.12598.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; de Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-Efficient Transfer Learning for NLP. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, 2790–2799.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

Li, Y.; Liu, H.; Wu, Q.; Mu, F.; Yang, J.; Gao, J.; Li, C.; and Lee, Y. J. 2023. GLIGEN: Open-Set Grounded Text-to-Image Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, 22511–22521.

Lin, T.; Maire, M.; Belongie, S. J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In Fleet, D. J.; Pajdla, T.; Schiele, B.; and Tuytelaars, T., eds., *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, 740–755.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Lu, C.; Zhou, Y.; Bao, F.; Chen, J.; Li, C.; and Zhu, J. 2022. DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Mou, C.; Wang, X.; Xie, L.; Wu, Y.; Zhang, J.; Qi, Z.; and Shan, Y. 2024. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. 38(5): 4296–4304.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P. F.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Park, T.; Liu, M.-Y.; Wang, T.-C.; and Zhu, J.-Y. 2019. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2337–2346.

Pernias, P.; Rampas, D.; and Aubreville, M. 2023. Wuerstchen: Efficient pretraining of text-to-image models.

Pfeiffer, J.; Kamath, A.; Rücklé, A.; Cho, K.; and Gurevych, I. 2021. AdapterFusion: Non-Destructive Task Composition for Transfer Learning. In Merlo, P.; Tiedemann, J.; and Tsarfaty, R., eds., *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, 487–503.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Qiu, Z.; Liu, W.; Feng, H.; Xue, Y.; Feng, Y.; Liu, Z.; Zhang, D.; Weller, A.; and Schölkopf, B. 2023. Controlling text-to-image diffusion by orthogonal finetuning. volume 36, 79320–79362.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *CoRR*, abs/2204.06125.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, 8821–8831. Pmlr.

Ranftl, R.; Lasinger, K.; Hafner, D.; Schindler, K.; and Koltun, V. 2022. Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(3): 1623–1637.

Richardson, E.; Alaluf, Y.; Patashnik, O.; Nitzan, Y.; Azar, Y.; Shapiro, S.; and Cohen-Or, D. 2021. Encoding in Style: A StyleGAN Encoder for Image-to-Image Translation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 2287–2296.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, 22500–22510.

Saharia, C.; Chan, W.; Chang, H.; Lee, C.; Ho, J.; Salimans, T.; Fleet, D.; and Norouzi, M. 2022a. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*, 1–10.

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. 2022b. Photorealistic text-to-image diffusion models with deep language understanding. volume 35, 36479–36494.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2256–2265. PMLR.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations.

Taigman, Y.; Polyak, A.; and Wolf, L. 2017. Unsupervised Cross-Domain Image Generation. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Wang, T.; Zhang, T.; Zhang, B.; Ouyang, H.; Chen, D.; Chen, Q.; and Wen, F. 2022. Pretraining is all you need for image-to-image translation. *arXiv preprint arXiv:2205.12952*.

Wang, T.-C.; Liu, M.-Y.; Zhu, J.-Y.; Tao, A.; Kautz, J.; and Catanzaro, B. 2018. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8798–8807.

Xie, S.; and Tu, Z. 2015. Holistically-Nested Edge Detection. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, 1395–1403.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, 3813–3824.

Zhao, S.; Chen, D.; Chen, Y.-C.; Bao, J.; Hao, S.; Yuan, L.; and Wong, K.-Y. K. 2024. Uni-controlnet: All-in-one control to text-to-image diffusion models. volume 36.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.