T2I-Adapter: Learning Adapters to Dig Out More Controllable Ability for Text-to-Image Diffusion Models

Chong Mou[*1,2], **Xintao Wang**[†2], **Liangbin Xie**[*2,3,4], **Yanze Wu**[2], **Jian Zhang**[†1], **Zhongang Qi**[2], **Ying Shan**[2]

[1]Peking University Shenzhen Graduate School
[2]ARC Lab, Tencent PCG
[3]University of Macau
[4]Shenzhen Institute of Advanced Technology
{eechongm, xintao.alpha, wuyanze123}@gmail.com, lb.xie@siat.ac.cn, zhangjian.sz@pku.edu.cn,
{zhongangqi, yingsshan}@tencent.com

Abstract

The incredible generative ability of large-scale text-to-image (T2I) models has demonstrated strong power of learning complex structures and meaningful semantics. However, relying solely on text prompts cannot fully take advantage of the knowledge learned by the model, especially when flexible and accurate controlling (*e.g.*, structure and color) is needed. In this paper, we aim to "dig out" the capabilities that T2I models have implicitly learned, and then explicitly use them to control the generation more granularly. Specifically, we propose to learn low-cost **T2I-Adapters** to align internal knowledge in T2I models with external control signals, while freezing the original large T2I models. In this way, we can train various adapters according to different conditions, achieving rich control and editing effects in the color and structure of the generation results. Further, the proposed T2I-Adapters have attractive properties of practical value, such as composability and generalization ability. Extensive experiments demonstrate that our T2I-Adapter has promising generation quality and a wide range of applications. Our code is available at https://github.com/TencentARC/T2I-Adapter.

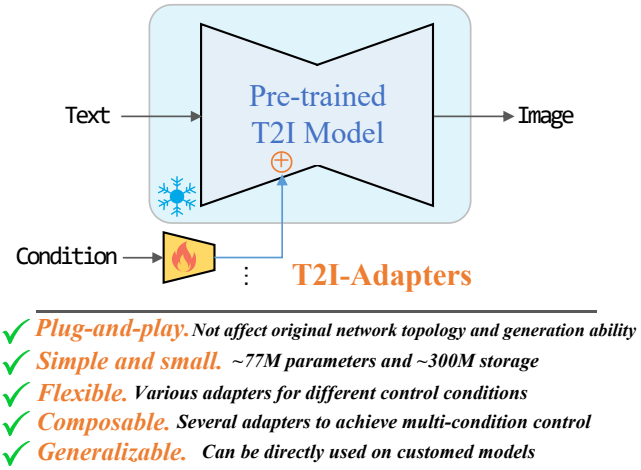

Figure 1: Our T2I-Adapter has several attractive properties to provide external guidance to pre-trained text-to-image models while not affecting their original generation ability.

Introduction

Thanks to the training on massive data and huge computing power, text-to-image (T2I) generation (Saharia et al. 2022; Rombach et al. 2022; Nichol et al. 2022; Ramesh et al. 2021; Ding et al. 2021; Zhou et al. 2021; Ramesh et al. 2022a; Gafni et al. 2022), which aims to generate images conditioned on a given text/prompt, has demonstrated strong generation ability. The generation results usually have rich textures, clear edges, reasonable structures, and meaningful semantics. This phenomenon potentially indicates that T2I models can actually capture information of different levels in an *implicit* way.

Although promising synthesis quality can be achieved, it heavily relies on well-designed prompts (Liu and Chilton 2022; Pavlichenko and Ustalov 2022), and the generation pipeline also lacks flexible user control capability that can guide the generated images to realize users' ideas accurately. For an unprofessional user, the generated results are usually uncontrolled and unstable. For example, the recently proposed Stable Diffusion (SD) (Rombach et al. 2022) can not perform well in some imaginative scenarios, *e.g.*, *"A car with flying wings"* and *"A banana and two apples on a plate"* as shown in Fig. 2. We believe that this does not mean that T2I models do not have the ability to generate such structures, just that the text cannot provide accurate structure guidance in random generation. In this paper, we are curious about whether it is possible to "dig out" the capabilities that T2I models have implicitly learned, especially the high-level structure and semantic capabilities, and then explicitly use them to control the generation more accurately.

Recently, some works provide guidance in T2I generation by efficient network tuning, such as Lora (Ryu 2023) inspired by rank decomposition (Hu et al. 2021) in NLP, DreamBooth (Ruiz et al. 2023) for generating specific char-

[*]Interns in ARC Lab, Tencent PCG
[†]Corresponding author

acters, Textual Inversion (Gal et al. 2022), etc. However, these works focus on fine-tuning the generation of specific objects/style and cannot provide structural control over the generated results. In this work, we achieve precise structure control over the generated results through efficient adapter tuning. We also noticed that the concurrent work ControlNet (Zhang and Agrawala 2023) studies this issue and achieves impressive results. In comparison, we regard external control as an alignment ability by several low-complexity adapters rather than a siamese network of the UNet encoder. Meanwhile, ControlNet, as a part of the denoiser, needs to participate in each diffusion step, while our T2I-Adapter only requires a single inference to inject the guiding information into each diffusion step. As shown in Fig. 1, the proposed T2I adapters have the following properties of practical value:

- **Plug-and-play**. They do not affect the original network topology and generation ability of existing T2I diffusion models (*e.g.*, Stable Diffusion).
- **Simple and small**. They can be easily inserted into existing T2I diffusion models with low costs ($\sim$ **77 M** parameters and $\sim$ **300 M** storage space), and they only need one inference during the diffusion process.
- **Flexible**. We can train various adapters for different external conditions, including spatial structure control and spatial color distribution of images.
- **Composable**. More than one adapter can be easily composed to achieve multi-condition control.
- **Generalizable**. Once trained, they can be directly used on custom models as long as they are fine-tuned from the same T2I model.

Our contributions are summarized as follows: **1).** We propose T2I-Adapter, a simple, efficient yet effective method to well align the internal knowledge of T2I models and external control signals at a low cost. **2).** T2I-Adapter can provide accurate controllable guidance to existing T2I models while not affecting their original generation ability. **3).** Extensive experiments demonstrate that our method works well with various conditions, and these conditions can also be easily composed to achieve multi-condition control.

Related Work

Text-to-image generation. Recently, autoregressive models (Gafni et al. 2022; Ramesh et al. 2021; Wu et al. 2022; Yu et al. 2022) and diffusion models (Nichol et al. 2022; Rombach et al. 2022; Ramesh et al. 2022b; Saharia et al. 2022) are dominant in the community of text-to-image (T2I) generation. Among autoregressive models, DALL-E (Ramesh et al. 2021) demonstrates the zero-shot T2I capability, make-a-scene (Gafni et al. 2022) presents attractive T2I generation quality. At the same time, abundant diffusion-based T2I methods are presented with promising performance. For instance, Glide (Nichol et al. 2022) proposes to combine the text feature into transformer blocks in the denoising process. Subsequently, DALL-E2 (Ramesh et al. 2022b), Stable Diffusion (Rombach et al. 2022) and Imagen (Saharia et al. 2022) vastly improve the performance in T2I generation. In particular, Stable Diffusion, which performs diffusion generation in the latent space, achieves state-of-the-art performance. Although they achieve promising synthesis quality, the text prompt can not provide the synthesis results with reliable structural guidance, resulting in highly random and uncontrollable results.

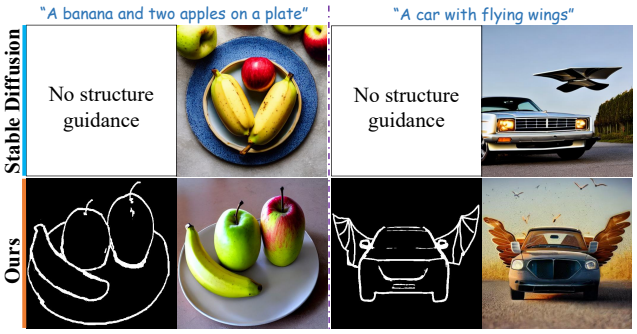

Figure 2: In some complex scenarios, SD (Rombach et al. 2022) fails to generate accurate results conforming to the text, as shown in the first row. In such cases, our T2I-Adapter can serve as a plugin to help SD generate reasonable results, as shown in the second row.

Conditional image generation. Conditional image generation aims to generate images with specific content by giving several relevant conditions. Most early works are based on generative adversarial networks (GAN) (Creswell et al. 2018), *e.g.*, (Isola et al. 2017; Park et al. 2019a; Wang et al. 2018; Zhu et al. 2017; Huang et al. 2022) propose generating natural images conditioned on specific condition maps in other domains (*e.g.*, sketch, semantic segmentation). Due to the significant improvement in generation quality and stability of diffusion models (Ho, Jain, and Abbeel 2020), most recent works focus on conditional image generation based on diffusion models. For instance, (Voynov, Aberman, and Cohen-Or 2022) guides image generation by using the similarity gradient produced by the target sketch and intermediate results. (Wang et al. 2022) proposes mapping the spatial structure control to the original text embedding of the T2I model (Nichol et al. 2022). Some methods (Song et al. 2022; Yang et al. 2023) introduce target object information into text tokens to achieve the insertion of specific objects in the generated results. (Cheng et al. 2023; Zheng et al. 2023; Li et al. 2023; Zeng et al. 2023) incorporate layout information into the diffusion generation process, enabling customized layouts in the generated results. As concurrent works, (Zhang and Agrawala 2023) learns task-specific ControlNet to enable conditional generation for the pre-trained T2I model. (Huang et al. 2023) proposes to retrain a new diffusion model conditioned on a set of control factors.

Efficient tuning on large models. Training large models is costly, and therefore, it is not efficient to fine-tune the entire large model for each downstream task. There has been extensive research on efficient fine-tuning of large language models (LLM), *e.g.*, (Houlsby et al. 2019) utilizes several adapters to transfer LLM to downstream tasks. LoRA (Hu

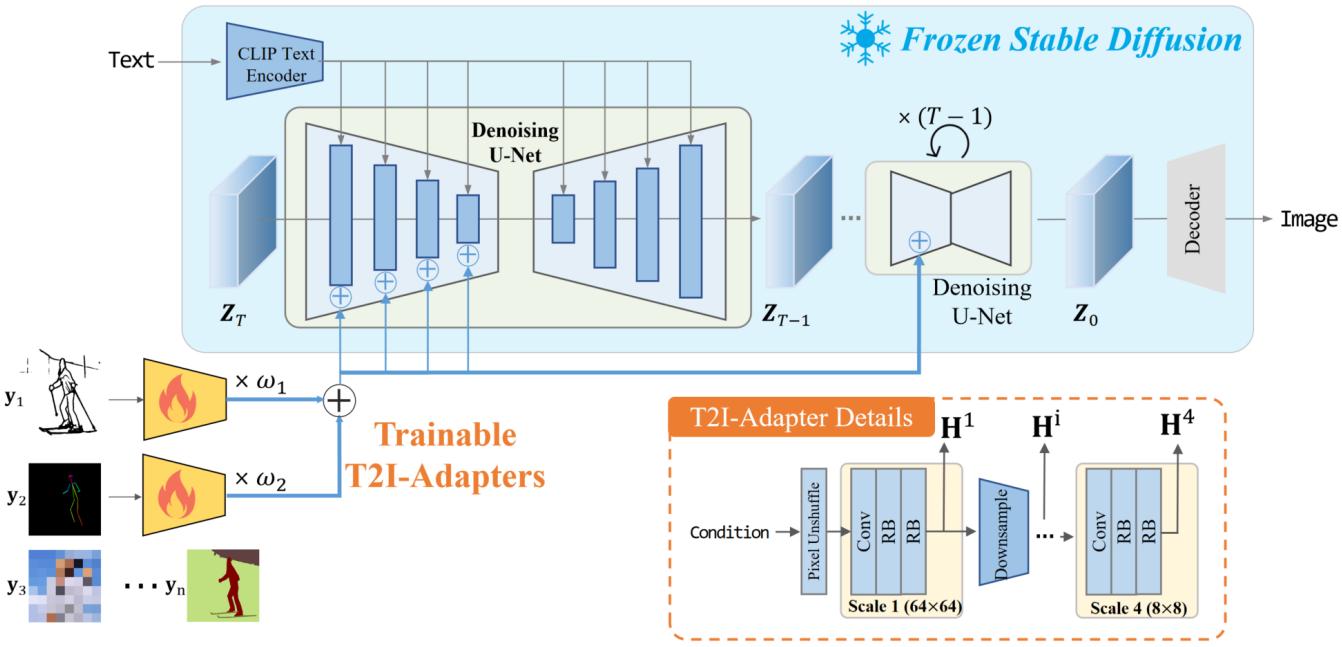

Figure 3: The overview of T2I-Adapter pipeline, which is composed of two parts: 1) a pre-trained stable diffusion with fixed parameters; 2) several T2I-Adapters trained to align internal knowledge in SD with external control signals. Different adapters can be composed by directly adding with adjustable weight ω.

et al. 2021) proposes to freeze the pre-trained LLM and inject trainable rank decomposition matrices into each layer of the Transformer architecture, greatly reducing the number of trainable parameters for downstream tasks. (Ryu 2023) applies the idea of LoRA to diffusion models, enabling the pre-trained Stable Diffusion (Rombach et al. 2022) (SD) to generate specific characters or styles. Subsequently, several efficient fine-tuning methods (Ruiz et al. 2023; Gal et al. 2022) for SD are proposed. In this paper, we insert several low-cost adapters to guide the generation of SD.

Method

Preliminary: Stable Diffusion

In this paper, we implement our method based on the recent T2I diffusion model (*i.e.*, Stable Diffusion (SD) (Rombach et al. 2022)). SD is a latent diffusion model (LDM), containing an autoencoder and an UNet denoiser. The autoencoder can convert the image $\mathbf{x}_0$ into latent space $\mathbf{z}_0$ and then reconstruct it. The diffusion process is performed in the latent space by a modified UNet denoiser. The optimization process can be defined as the following formulation:

$$\mathcal{L} = \mathbb{E}_{\mathbf{z}_t,\mathbf{C},\epsilon,t}(||\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \mathbf{C})||_2^2), \quad (1)$$

where $\mathbf{z}_t$ represents the noised latent at time step t. $\mathbf{C}$ represents the conditional text embedding generated by the pre-trained CLIP (Radford et al. 2021) text encoder. ϵ_θ refers to the function of UNet denoiser. During sampling, the latent $\mathbf{z}_t$ are gradually denoised from the initial random Gaussian noise through ϵ_θ conditioned on $\mathbf{C}$ and t. Finally, the denoised latent is converted to an image by the decoder of the autoencoder.

Overview of T2I-Adapter

As shown in the first row of Fig. 2, the text can hardly provide structural guidance to image synthesis, leading to random and unstable composition of generated results in terms of spatial structure. Based on this observation, we want to provide customized spatial alignment for the generation of SD, which cannot be provided by text alone. We believe that the alignment should not be considered as a new generation capability for large T2I models to relearn, but rather as a capability that can be easily learned through external plugins. An overview of our method is presented in Fig. 3, which is composed of a pre-trained SD model and several T2I adapters. The adapters are used to extract guidance features from different types of conditions, *e.g.*, sketch, canny, keypoints, color, depth, and semantic segmentation. The pre-trained SD has fixed parameters to generate images based on the input text feature and external guidance feature. As can be seen from our pipeline, all trainable parameters are in the additional adapters, which can be removed at any time without affecting the original T2I model.

Adapter Design

Our proposed T2I-Adapter is simple and lightweight, as shown in the right corner of Fig. 3. It is composed of four feature extraction blocks and three downsample blocks to change the feature resolution. The condition input has the resolution of $w \times h$. Here, we utilize the pixel unshuffle (Shi et al. 2016) operation to downsample it to $\frac{w}{8} \times \frac{h}{8}$. In each scale, one convolution layer and two residual blocks (RB) are utilized to extract the condition feature $\mathbf{H}^i$, $i = 1, 2, 3, 4$. Finally, multi-scale condition fea-

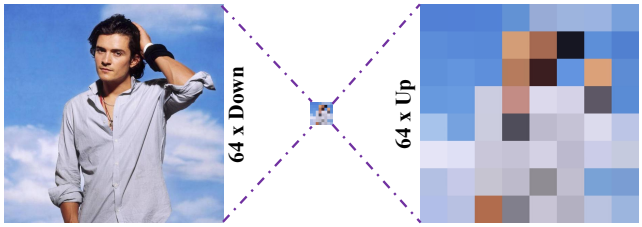

Figure 4: Illustration of spatial color condition. We erase image details by extreme downsampling while retaining the approximate color information and its spatial distribution.

tures $\mathbf{H} = \{\mathbf{H}^1, \mathbf{H}^2, \mathbf{H}^3, \mathbf{H}^4\}$ are generated. Note that the dimension of $\mathbf{H}$ is the same as the intermediate feature $\mathbf{F}_{enc} = \{\mathbf{F}^1_{enc}, \mathbf{F}^2_{enc}, \mathbf{F}^3_{enc}, \mathbf{F}^4_{enc}\}$ in the encoder of UNet denoiser. $\mathbf{H}$ is then **added** with $\mathbf{F}_{enc}$ at each scale. Mathematically, the condition feature extraction and condition injection are formulated as:

$$\mathbf{H} = \mathcal{A}(\mathbf{y}) \tag{2}$$

$$\hat{\mathbf{F}}^i_{enc} = \mathbf{F}^i_{enc} + \mathbf{H}^i,\ i \in \{1, 2, 3, 4\}, \tag{3}$$

where $\mathbf{y}$ is the condition input. $\mathcal{A}$ is function of T2I-Adapter.

Structure controlling. Our T2I-Adapter demonstrates strong generalization capabilities, supporting a wide range of structure controls such as sketch, canny, depth, semantic segmentation, and keypoint. These controls are obtained through specific operators, with details provided in the experiment section. The condition maps are directly input into task-specific adapters to extract condition features $\mathbf{H}$. These adapters share the same structure as described above.

Spatial color palette. In addition to spatial structure, color is also an important component of images. Similarly, color has spatial attributes, defining the color at different positions. In this paper, we design a spatial color palette to roughly control the color distribution of the generated images. The representation of this condition is shown in Fig. 4. Specifically, we use aggressive ($\times 64$) bicubic downsampling to remove the semantic and structural information of the image while preserving enough color information. Then we apply the nearest upsampling to restore the original size of the image. Finally, the spatial color condition is represented by several spatial-arrangement color blocks, and the guidance information is injected into the diffusion process in the same way as other structure conditions.

Multi-adapter controlling. In addition to using a single adapter to guide the generation, our T2I-Adapter also supports multi-condition control. It is completed through weighted sum without additional training, and the control strength of different conditions can be adjusted by weights. Mathematically, this process is defined as:

$$\mathbf{H} = \sum_{n=1}^{N} \omega_n \mathcal{A}_n(\mathbf{y}_n), \tag{4}$$

where N represents the number of conditions. ω_n is the adjustable weight to control the strength of each adapter. This composable property leads to several useful applications.

Figure 5: We evenly divide the DDIM inference sampling into 3 stages, *i.e.*, beginning, middle and late stages. We show the results of adding guidance at these three stages.

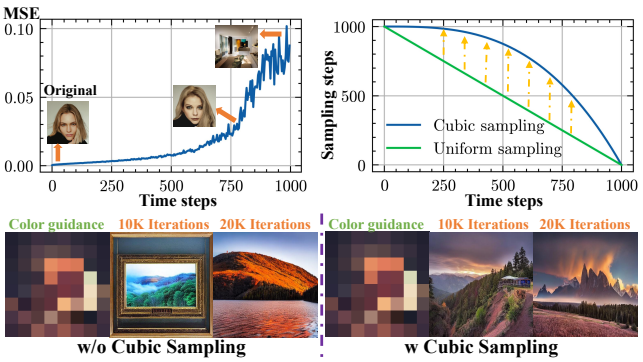

Figure 6: The first curve represents the loss (MSE) caused by adding noise and then denoising at different time steps on the same image. The second curve is the mapping between uniform sampling and cubic sampling. The second row visualizes the effectiveness of cubic sampling in adapter training.

For instance, we can use the sketch map to provide structure guidance while using the spatial color palette to color the results. An example is presented in Fig. 11.

Adapter Training

During training, we fix the parameters in SD and only optimize the T2I-Adapter $\mathcal{A}$. Each training sample is a triplet, including the original image $\mathbf{x}_0$, condition map $\mathbf{y}$, and text embedding $\mathbf{C}$. The optimization process is similar to SD. Specifically, given an image $\mathbf{x}_0$, we first embed it to the latent space $\mathbf{z}_0$ via the encoder of autoencoder. Then, we randomly sample a time step t from $[0, T]$ and add corresponding noise to $\mathbf{z}_0$, producing $\mathbf{z}_t$. Mathematically, our T2I-Adapter is optimized via:

$$\mathcal{L} = \mathbb{E}_{\mathbf{z}_0, t, \mathbf{H}, \epsilon \sim \mathcal{N}(0,1)} \left[||\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \mathbf{C}, \mathcal{A}(\mathbf{y}))||_2^2 \right] \tag{5}$$

Non-uniform time step sampling during training. In diffusion models, the time step t is an important condition, indicating the noise intensity at each time step. In this paper, we study its role in training low-cost adapters and design a non-uniform sampling strategy to improve adapter training.

There is an observation, shown in Fig. 5. Specifically, we evenly divide the 50-step DDIM sampling into 3 stages, *i.e.*, beginning, middle and late stages. We then add guidance information to each of the three stages. We find that adding guidance in the middle and late stages have little effect on the result. It indicates that the main content of the generation results is determined in the early sampling stage. To

	Sketch	Canny	Depth	Keypoint	Semantic segmentation
SPADE	-/-	-/-	-/-	-/-	23.44/0.2314
OASIS	-/-	-/-	-/-	-/-	18.71/0.2274
PITI	21.21/0.2129	-/-	-/-	-/-	17.36/0.2287
GLIGEN	-/-	19.01/0.2520	21.05/0.2609	32.41/0.2496	23.79/0.2490
ControlNet	19.84/**0.2638**	**15.73/0.2613**	19.09/0.2631	**28.93/0.2640**	18.78/**0.2653**
T2I-Adapter (Ours)	**18.30**/0.2593	17.96/0.2608	**18.14/0.2656**	29.77/0.2617	**16.78**/0.2652

Table 1: Quantitative comparison (FID↓ / CLIP Score↑ (ViT-L/14)) on COCO (Lin et al. 2014) validation set between our T2I-Adapter and other methods. ControlNet and T2I-Adapter employ 20 DDIM steps for fast evaluation.

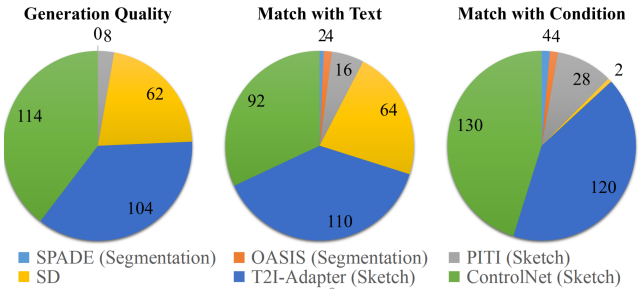

Figure 7: User study on generation quality and alignment accuracy of different methods (condition is labeled in bracket).

further verify it, we added noise with different time steps $t \in [0, 1000]$ to $\mathbf{z}_0$, producing $\mathbf{z}_t$. Then we use SD for denoising generation starting from $\mathbf{z}_t$. Note that a larger t indicates stronger noise and is closer to the beginning stage. We calculated the mean square error (MSE) between the generated result and the original image, as shown in the first curve in Fig. 6. It can be seen that adding noise in the middle and later stages has little impact on the final generated result. Only under the influence of high-intensity noise in the beginning stage, the generated result will show a larger deviation. Therefore, if too many time steps are sampled in the middle and later stages during training, the external guidance can be easily ignored by the network, because the noisy latent has enough information to reconstruct the original image.

To fully train the adapter, we adopt cubic time step sampling (*i.e.*, $t = (1 - (\frac{t}{T})^3) \times T$), as shown in the second curve in Fig. 6. It allows more time steps to be sampled in the high-intensity noise region, enhancing the role of external guidance during training. The importance of this sampling strategy is more evident in color control, as color, being low-level visual information, is more difficult to erase than structural information. The second row in Fig. 6 shows that the spatial color adapter has weak color control without cubic sampling strategy. After using cubic sampling, the spatial color adapter can converge rapidly and well control the color distribution.

Experiment

Implementation Details

We choose the pre-trained SD-V1.5 (Rombach et al. 2022) as the base model. During training, we utilize Adam (Kingma and Ba 2014) as the optimizer with the learning rate of 1×10^{-5}. The input images and condition maps are resized to 512×512. The training process is performed on 4 NVIDIA Tesla 32G-V100 GPUs with a batch size of 8, which can be completed within 3 days.

Training data. Experiment includes 6 types of conditions:

- *Semantic segmentation.* In this application, we utilize COCO-Stuff (Caesar, Uijlings, and Ferrari 2018) as the training data, which contains $164K$ images. Its semantic segmentation contains 80 thing classes, 91 stuff classes and 1 *'unlabeled'* class.
- *Sketch & Canny & Color & Depth.* For these applications, we use images from LAION-AESTHETICS (Schuhmann et al. 2022) dataset as the training data. The sketch and canny are obtained through edge detection algorithms (Su et al. 2021; Xu, Baojie, and Guoxin 2017), and the depth is obtained through MiDaS (Ranftl et al. 2022).
- *Keypoint.* For keypoint, we use openpose (Cao et al. 2019) to extract the keypoint map of each image from LAION-AESTHETICS dataset and finally select about $500K$ images containing human bodies.

Benchmark and metrics. To validate the performance of different methods, we choose COCO validation set (Lin et al. 2014) as the test set, which contains 5,000 image-text pairs. Since each image contains 5 text descriptions, we use the first text of each group as input during testing. To quantify performance, we use user study, FID (Seitzer 2020), and CLIP score (Radford et al. 2021) as evaluation metrics. The computational complexity is also considered.

Comparison

In this part, we compare our method with some GAN-based methods (*i.e.*, SPADE (Park et al. 2019b), OASIS (Schönfeld et al. 2021)) and diffusion-based methods (*i.e.*, PITI (Wang et al. 2022), GLIGEN (Li et al. 2023) and ControlNet (Zhang and Agrawala 2023)). The quantitative comparison in Tab. 1 shows that the performance of our method is comparable to ControlNet under 5 conditions and is superior to other methods. We also conduct a user study in Fig. 7, including three aspects: (1) generation quality; (2) matching with text; (3) matching with condition. Participants should select the best one for each metric from each comparison group. We collect votes from 24 participants with 12 cases. One can see that SD is indeed

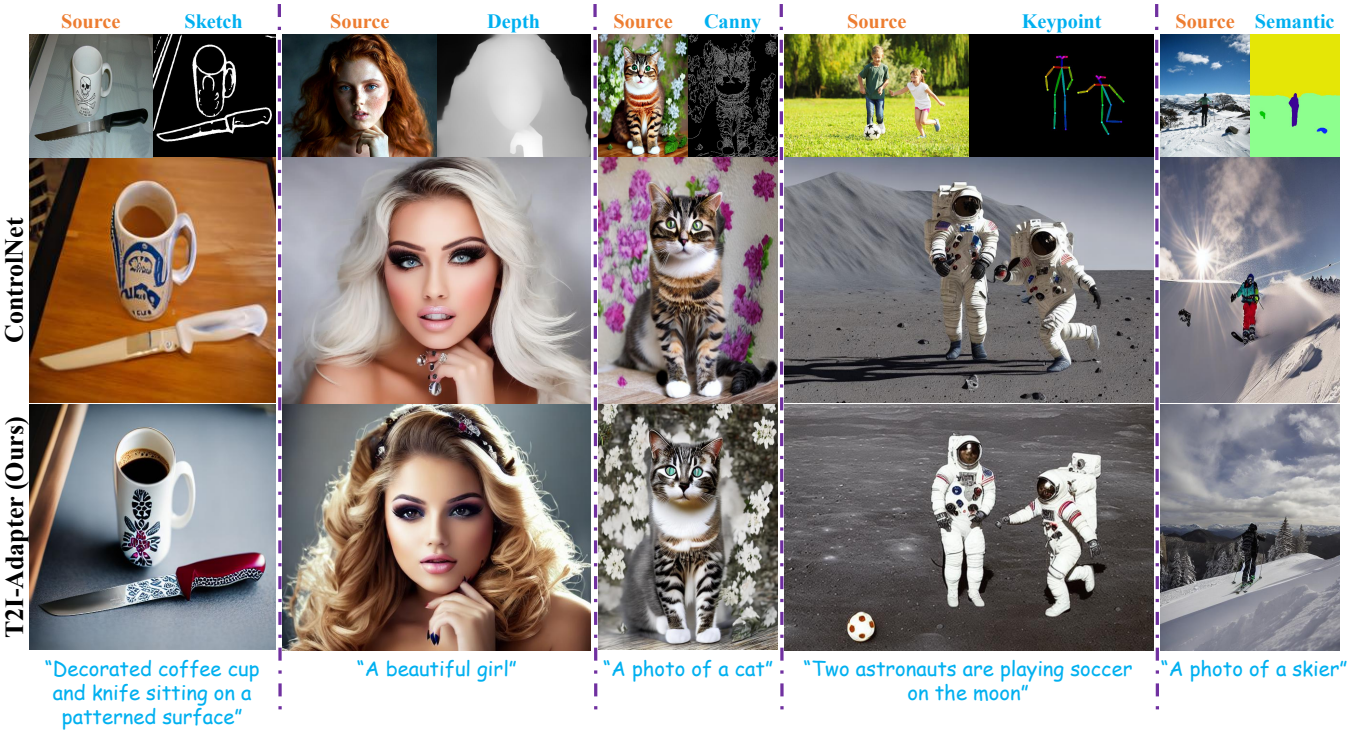

Figure 8: The visual comparison between our T2I-Adapter and ControlNet (Zhang and Agrawala 2023) under 5 control conditions, *i.e.*, sketch, depth, canny, keypoint, and semantic segmentation.

a good text-based generation prior, leading ControlNet and our T2I-Adapter to be significantly better than other methods in terms of generation quality, text alignment and condition alignment. Compared with ControlNet, our method can achieve comparable performance with a smaller cost.

In Fig. 8, we show visual comparisons between our T2I-Adapter and ControlNet under 5 conditions on SD-V1.5. We present the complexity comparison in Tab. 2. Note that the inference speed is measured by generating 512x512 images on a single NVIDIA A100 GPU. Both our method and ControlNet use XFormers (Lefaudeux et al. 2022) for acceleration. We use the number of diffusion iterations per second (it/s) to represent the inference speed. We can find that our method has comparable performance to ControlNet, while the model parameters and inference speed are $\frac{1}{7}$ and 1.5 times that of ControlNet, respectively. In addition, we compare our T2I-Adapter and ControlNet on larger SD *i.e.*, SDXL (Podell et al. 2023), which has 2.6B parameters. On SDXL, we use the T2I-Adapter with similar structures and parameters to that in SD-V1.5. The visual and complexity comparison are presented in Fig. 9 and Tab. 2, respectively. One can see that the performance of **79M** T2I-Adapter-XL and **1250M** ControlNet-SDXL has close performance. Therefore, the gains from larger control models are limited and do not continue to increase with model size.

Other Capabilities

Spatial color palette. Our method also supports spatial color control, as shown in Fig. 10. The input spatial color grid can effectively control the color of the generated result in different regions.

	Ctrl	T2I	Ctrl-XL	T2I-XL
Param.↓	567M	**77M**	1250M	**79M**
Speed(it/s)↑	18.78	**27.96**	10.85	**6.40**

Table 2: Complexity comparison between our T2I-Adapter (T2I) and ControlNet (Ctrl) on SD and SDXL.

Figure 9: The performance and complexity comparison between T2I-Adapter and ControlNet on **SDXL**.

Composability of different conditions. We find that adapters for each condition can be well combined after being trained separately, and this ability does not require additional training. The combination is completed by the weighted sum of different condition features as shown in Eq. 4. In Fig. 11, we present the generation results of T2I-Adapter under multiple conditions. It can be seen that the input of the four con-

Figure 10: Results of T2I-Adapter on spatial color control.

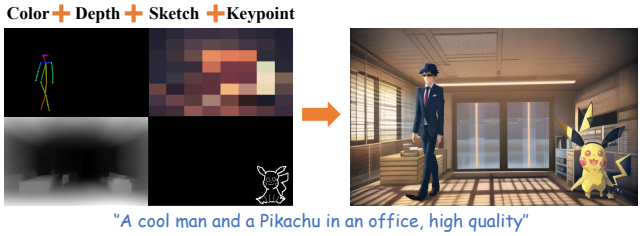

Figure 11: Results of our T2I-Adapter under the control of multiple conditions, *i.e.*, keypoint, color, depth and sketch.

ditions achieves their respective generation goals, and these generated contents are naturally blended together.

Generalization to other fine-tuned SD models. The generalization ability of the adapter is an interesting and useful property. Concretely, once adapters are trained, they can be directly used on custom models as long as they are fine-tuned from the same T2I model. As shown in Fig. 12, we download custom models fine-tuned on SD from https://civitai.com/ and then directly insert our T2I-Adapter into them. One can see that even without specific training, our method still achieves good control effects. This generalization ability allows our T2I-Adapter to have a wider range of applications after a single training.

Ablation Study

In this part, we conduct ablation study on the guidance accuracy by adding conditions at different positions of the UNet denoiser. The experiment is conducted on the COCO validation set with sketch guidance. To measure the matching between the generated results and the condition, we extract the sketch from the generated results and calculate its similarity with the input sketch using the mean squared error (MSE). The smaller the value, the higher the matching.

The result is presented in Tab. 3. Specifically, we evaluate the performance of adding guidance information to the encoder and decoder of the UNet denoiser. The result presents that our T2I-Adapter has better control capabilities when guidance information is added to the encoder. This is mainly because the control information can flow through a longer network pipeline (*i.e.*, encoder and decoder), allowing the guidance information to be fully integrated with the intermediate features. It serves as a performance compensation for our low-complexity adapters. Adding guidance to both the encoder and decoder leads to better performance, but the complexity will double. In addition, we also conducted ablation study on the four scales of the UNet encoder, and it can be seen that the multi-scale control injection is beneficial for improving control accuracy. Finally, considering both performance and computational complexity, we choose to add guidance information to the UNet encoder with four scales.

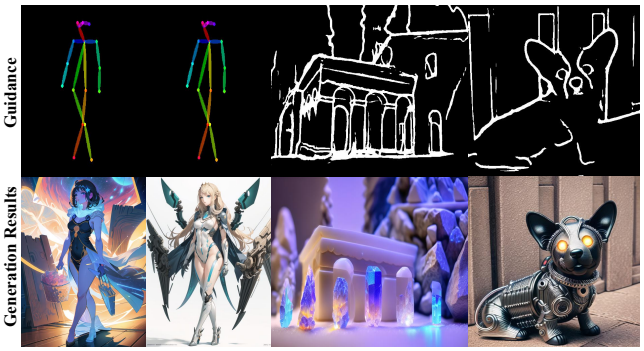

Figure 12: The controlling performance of our T2I-Adapter on other custom models without additional training.

Mode	Scale Num.	Enc.	Dec.	MSE.↓
1	4	✓	✗	0.1280
2	4	✗	✓	0.1455
3	4	✓	✓	0.1207
4	3	✓	✗	0.1427
5	2	✓	✗	0.1878
6	1	✓	✗	0.2023

Table 3: Ablation study of how the guidance information is injected into the SD model.

Conclusion and Limitation

In this paper, we aim to dig out the capabilities that T2I models have implicitly learned, *e.g.*, the colorization and structuring capabilities, and then explicitly use them to control the generation more accurately. We present that a low-cost adapter model can achieve this purpose, as it is not learning new generation abilities but learning an alignment between external control signals and internal knowledge in pre-trained T2I models. Even in larger diffusion models, *e.g.*, SDXL, our method can drive it effectively. In addition to the simplicity and lightweight structure, our T2I-Adapter 1) does not affect the original generation ability of the pre-trained T2I model; 2) has a wide range of applications in spatial color control and elaborate structure control. 3) More than one adapter can be easily composed to achieve multi-condition control. 4) Once trained, the T2I-Adapter can be directly used on custom models as long as they are fine-tuned from the same T2I model. Finally, extensive experiments demonstrate that the proposed T2I-Adapter achieves excellent controlling and promising generation quality. One limitation of our method is that in the case of multi-adapter control, the combination of guidance features requires manual adjustment. In our future work, we will explore the adaptive fusion of multi-modal guidance information.

Acknowledgements

This work was supported by National Natural Science Foundation of China under Grant 62372016.

References

Caesar, H.; Uijlings, J.; and Ferrari, V. 2018. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1209–1218.

Cao, Z.; Hidalgo Martinez, G.; Simon, T.; Wei, S.; and Sheikh, Y. A. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Cheng, J.; Liang, X.; Shi, X.; He, T.; Xiao, T.; and Li, M. 2023. Layoutdiffuse: Adapting foundational diffusion models for layout-to-image generation. *arXiv preprint arXiv:2302.08908*.

Creswell, A.; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B.; and Bharath, A. A. 2018. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1): 53–65.

Ding, M.; Yang, Z.; Hong, W.; Zheng, W.; Zhou, C.; Yin, D.; Lin, J.; Zou, X.; Shao, Z.; Yang, H.; et al. 2021. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34: 19822–19835.

Gafni, O.; Polyak, A.; Ashual, O.; Sheynin, S.; Parikh, D.; and Taigman, Y. 2022. Make-a-scene: Scene-based text-to-image generation with human priors. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, 89–106. Springer.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

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 *International Conference on Machine Learning*, 2790–2799. PMLR.

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. *arXiv preprint arXiv:2106.09685*.

Huang, L.; Chen, D.; Liu, Y.; Yujun, S.; Zhao, D.; and Jingren, Z. 2023. Composer: Creative and Controllable Image Synthesis with Composable Conditions. *arXiv preprint arxiv:2302.09778*.

Huang, X.; Mallya, A.; Wang, T.-C.; and Liu, M.-Y. 2022. Multimodal conditional image synthesis with product-of-experts gans. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI*, 91–109. Springer.

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.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lefaudeux, B.; Massa, F.; Liskovich, D.; Xiong, W.; Caggiano, V.; Naren, S.; Xu, M.; Hu, J.; Tintore, M.; Zhang, S.; Labatut, P.; and Haziza, D. 2022. xFormers: A modular and hackable Transformer modelling library. https://github.com/facebookresearch/xformers.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22511–22521.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, V.; and Chilton, L. B. 2022. Design guidelines for prompt engineering text-to-image generative models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–23.

Nichol, A. Q.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; Mcgrew, B.; Sutskever, I.; and Chen, M. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *International Conference on Machine Learning*, 16784–16804. PMLR.

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

Park, T.; Liu, M.-Y.; Wang, T.-C.; and Zhu, J.-Y. 2019b. Semantic Image Synthesis with Spatially-Adaptive Normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Pavlichenko, N.; and Ustalov, D. 2022. Best Prompts for Text-to-Image Models and How to Find Them. *arXiv preprint arXiv:2209.11711*.

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*.

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.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022a. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022b. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv: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 Transactions on Pattern Analysis and Machine Intelligence*, 44(3).

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22500–22510.

Ryu, S. 2023. Low-rank adaptation for fast text-to-image diffusion fine-tuning.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.; Ghasemipour, S. K. S.; Ayan, B. K.; Mahdavi, S. S.; Lopes, R. G.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*.

Schönfeld, E.; Sushko, V.; Zhang, D.; Gall, J.; Schiele, B.; and Khoreva, A. 2021. You Only Need Adversarial Supervision for Semantic Image Synthesis. In *International Conference on Learning Representations*.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294.

Seitzer, M. 2020. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid. Version 0.3.0.

Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A. P.; Bishop, R.; Rueckert, D.; and Wang, Z. 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1874–1883.

Song, Y.; Zhang, Z.; Lin, Z.; Cohen, S.; Price, B.; Zhang, J.; Kim, S. Y.; and Aliaga, D. 2022. Objectstitch: Generative object compositing. *arXiv preprint arXiv:2212.00932*.

Su, Z.; Liu, W.; Yu, Z.; Hu, D.; Liao, Q.; Tian, Q.; Pietikäinen, M.; and Liu, L. 2021. Pixel difference networks for efficient edge detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5117–5127.

Voynov, A.; Aberman, K.; and Cohen-Or, D. 2022. Sketch-Guided Text-to-Image Diffusion Models. *arXiv preprint arXiv:2211.13752*.

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.

Wu, C.; Liang, J.; Ji, L.; Yang, F.; Fang, Y.; Jiang, D.; and Duan, N. 2022. Nüwa: Visual synthesis pre-training for neural visual world creation. In *European conference on computer vision*, 720–736. Springer.

Xu, Z.; Baojie, X.; and Guoxin, W. 2017. Canny edge detection based on Open CV. In *2017 13th IEEE international conference on electronic measurement & instruments (ICEMI)*, 53–56. IEEE.

Yang, B.; Gu, S.; Zhang, B.; Zhang, T.; Chen, X.; Sun, X.; Chen, D.; and Wen, F. 2023. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18381–18391.

Yu, J.; Xu, Y.; Koh, J. Y.; Luong, T.; Baid, G.; Wang, Z.; Vasudevan, V.; Ku, A.; Yang, Y.; Ayan, B. K.; et al. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3): 5.

Zeng, Y.; Lin, Z.; Zhang, J.; Liu, Q.; Collomosse, J.; Kuen, J.; and Patel, V. M. 2023. Scenecomposer: Any-level semantic image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22468–22478.

Zhang, L.; and Agrawala, M. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. arXiv:2302.05543.

Zheng, G.; Zhou, X.; Li, X.; Qi, Z.; Shan, Y.; and Li, X. 2023. LayoutDiffusion: Controllable Diffusion Model for Layout-to-image Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22490–22499.

Zhou, Y.; Zhang, R.; Chen, C.; Li, C.; Tensmeyer, C.; Yu, T.; Gu, J.; Xu, J.; and Sun, T. 2021. Lafite: Towards language-free training for text-to-image generation. *arXiv preprint arXiv:2111.13792*.

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.