A Survey of Personalizable Text to Image Diffusion Models

Andrew Zhong
University of Maryland

Abstract

Diffusion models have revolutionized text-to-image (T2I) generation with their ability to generate a wide variety of high-quality images. While T2I diffusion models are capable of fine-tuned control, text alone cannot capture task-specific requirements, prompting many studies on how to generate images based on user-provided input examples. This paper surveys the field of personalizable T2I diffusion models, covering both existing advancements and directions for future work. We begin with an overview of the theoretical basis of diffusion models and methods for conditioning image generation based on novel concepts. We then provide a detailed survey of advancements, organized into the following categories of conditions: generic concepts, people, interactions, and layouts. Finally, we discuss outstanding problems with existing models and future directions for work.

1 Introduction

In recent years, diffusion models have revolutionized the field of image generation [1]. Due to their stability and quality [2], they have displaced Generative Adversarial Networks (GANs) [3] as the prevailing base architecture in state-of-the-art image generation. While transformer-based methods such as Muse [4] are a promising new architecture with efficiency advantages, we will consider them out of scope for this paper. Text to image (T2I) diffusion models tackle the task of image generation from text prompts, a common use case with applications in digital art, video generation, and image editing.

As text is limited in the level of detail that it can provide, various methods have been introduced to control the generated output, such as custom subjects, styles, spatial layouts, and other concepts. This is imperative in applications where a unique concept has either not been captured during the training process due to not being seen before or cannot be described completely by text. Common examples include image generation with a specific person, object, style, action, or spatial layout [5][6][7][8][9][10]. In this survey, we will focus on the subdomain of controllable T2I diffusion models focusing on customizable text to image generation. Prior surveys have covered diffusion models as a whole [11][12][13], diffusion models for vision [14], video generation diffusion models [15], reinforcement learning diffusion models [16], and diffusion models for image editing [17]. Cao et al [5] survey controllable T2I diffusion models as a whole with a focus on summarizing the contributions of existing papers. In contrast, this paper focuses on a subcategory of controllable T2I models and provides analysis of the outstanding problems in the domain, highlighting avenues for future research.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

2 Background

2.1 Denoising Diffusion Probabilistic Models

Diffusion models, or Denoising Diffusion Probabilistic Models (DDPMs), are parameterized Markov chains which are trained to iteratively reverse a noise-adding process which iteratively destroys a signal[1]. By doing so, diffusion models learn a mapping from Gaussian noise to a target distribution. Image generation models utilize this to map random noise samples to realistic image distributions captured by large, internet-scale training datasets.

2.2 Theory

First, we will begin with the noising diffusion process. Given samples x_0 from a data distribution described by the probability density function $q(x_0)$, gaussian noise is incrementally added over T iterations until the signal is completely destroyed and only noise remains. This is called the forward process.

Each step of the forward noising process can be described by the following conditional probability. The variance hyperparameters of the added noise are β_t.

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}) \tag{1}$$

Accumulated over T timesteps, this is equivalent to:

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t, x_{t-1}) \tag{2}$$

In order to sample the desired distribution, the reverse process is approximated by a deep neural network, often based on UNet. Given a noise sample x_T the DDPM parameterizes each step as the following normal distribution, where θ denotes values predicted by the model.

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{3}$$

Given the sample x_t at time t, the neural network then predicts the noise ϵ_θ that needs to be added to x_t to recover x_{t-1}. After T iterations, the model is able to recover x_0 from the noise sample x_T. During training, all x_t are obtained from the corresponding training image x_0 through the forward diffusion process. To generate novel images, the noise vector x_T is sampled from the Gaussian distribution and passed through the reverse process to generate a new image. The goal is to obtain a new sample from the training distribution of real images by using our DDPM as a mapping.

2.3 Controlling Diffusion Models

The generated images can be controlled to follow text conditions and custom conditions. From a score-based perspective[5][18][19], a approximation s_θ for the following score function is incorporated into μ_θ.

$$\nabla_{x_t} log\left(p_t(x)p_t^w(x|c_{text}, c_{cond})\right) \tag{4}$$

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}x_t - \frac{1-\alpha_t}{\sqrt{\bar{\alpha}_t}}s_\theta(x_t, t) \tag{5}$$

where w is the weight of the conditioning, while c_{text} is the text condition and c_{cond} is the new condition that a given model is adding.

Given these conditions, we now predict our noise based on a sum weighted by w of noise predicted with and without conditions c_{text} and c_{cond}.

$$\epsilon(x_t, c_{text}, c_{cond}, t) = (1-w)\bar{\epsilon}(x_t, \phi, t) + w\bar{\epsilon}(x_t, c_{text}, c_{cond}, t) \tag{6}$$

The following three sections describe methods of predicting $\bar{\epsilon}(x_t, c_{text}, c_{cond}, t)$.

2.3.1 Fine-tuning Methods

In tuning-based methods, models are fine-tuned for a specific condition on small training sets of examples. Instead of providing the condition as an input, the condition is instead baked into the model itself, turning $\bar{\epsilon}(x_t, c_{text}, c_{cond}, t)$ into a prediction of $\bar{\epsilon}(x_t, c_{text}, t)$ instead.

2.3.2 Embedding-based Methods

Embedding-based methods utilize encoder models to extract features from conditions and pass them as input to ϵ_θ. As a result, we predict $\bar{\epsilon}(x_t, c_{text}, c_{cond}, t)$ as $\epsilon_\theta(x_t, c_{text}, e_{cond}, t)$, where e_{cond} is the encoding of c_{cond} by a trained encoder E.

2.3.3 Training-free Methods

Some methods do not require training at all and instead control general by features of the existing architecture. For example, many models utilize the cross attention layer to refocus attention maps according to a goal layout [20][21][22]. Techniques of this type have the incredible advantage of coming "for free" and are often usable across different base diffusion models.

3 Custom Concept Types

In this section, we will cover advancements and contributions of papers in dealing with the following concept types: subjects, styles, interactions, layouts, and more generalizable concepts. Additionally, we will offer commentary on limitations of varying techniques and compare the advantages/disadvantages of different models.

3.1 Generic Concepts

In this section, we will cover methods of capturing generic custom concepts or images. These models are often used as the basis for other methods.

Textual Inversion[8] utilizes the pre-existing representations for text token embeddings by training to capture concepts with a new multimodal prompt token. Rather than expanding the tokenizer's vocabulary, DreamBooth [7] trains to to overwrite a rare token (such as "sks") to represent an identifier for a unique subject. In addition, a class-specific preservation loss is used to ensure that learning a unique instance of a subject does not jeopardize the model's ability to generate diverse images that class. Without this learning a specific instance (such as an "sks cat") could cause the model to generate only that instance even when the prompt is the more general "cat". Custom Diffusion [6] greatly optimizes tuning time by updating only the most important weights. By analyzing the rate of change of weights during training, the authors were able to identify that most of the change during fine-tuning occurs in the cross attention layer, and propose to only update the W^k and W^v parameters of that layer.

Unlike Textual Inversion, which represents images directly with a token, DreamBooth elects to overwrite an existing token as a *modifier* for an existing class. While this allows for a greater degree of specificity as it identifies what part of the image is being referred to, it limits the generality of the model to capture unseen concepts. Custom Diffusion has both modifiers and can capture new concepts. However, in constrast to Textual Inversion where the token directly represents the image, Custom Diffusion includes the token as part of the text captions for the training examples. This allows for more detailed descriptions of the custom concept and what part of the reference image it belongs to.

Similarly to Textual Inversion, UnCLIP [23] utilizes CLIP [24] embeddings, which provide a unified representation for images and their text captions, to generate variations of custom images with no training. Unlike the aforementioned models, UnCLIP focuses more on creating variations of a custom image rather than new, text-controllable images containing a subject from the reference image. Similarly, Prompt-free Diffusion [25] aims to remove the "burden" of prompt engineering on the user's part by implementing a direct image to image pipeline controlled only by optional structural guidance.

Figure 1: Action customization using Action Disentangled Identifiers[34]. The actions/poses from example images (left) are learned and applied to novel subjects. Figure sourced from ADI paper[34].

An paper of note is InstructPix2Pix [26]. Instead of extracting a custom concept from reference images, InstructPix2Pix edits the (one) reference image itself following a text prompt. Amazingly, the model is trained on a generated dataset. Given a "before" text caption, GPT-3 [27] is used to generate editing instructions and "after" text captions, and Stable Diffusion along with Prompt2Prompt[28] are used to generate images corresponding to the before/after text captions.

3.2 People

Like many vision and image based models, there exist several human-specific models due to how common human-specific tasks are. For example, an emerging use case is virtual garment try-on, where given a person and a garment, the task is to generate an image where the person is wearing that garment in an identity preserving manner. [29][30]. Similarly, the major challenge in human-based custom diffusion models is preserving the identity and other characteristics of the subject.

Photoverse [31] introduces a novel facial identity loss to enforce identity preservation, resulting in a fine-tuning free model capable of generating diverse images from a single reference image. Multiple methods such as DreamIdentity [32] and Face0 [33] utilize face-identity encoders to create text embedding space representations for identity. These models are able to create diverse, identity-preserving images with zero fine-tuning on reference images. DreamIdentity also implements a novel training regime called "self-augmented editability learning" to leverage existing models' high performance on celebrity faces. Identity preserving images are generated with celebrity images, which are then used to train diffusion models that generalize to unseen faces.

3.3 Interactions

Interactions describe the actions, relationships, and other modifiers of subjects in an image. Huang et al[34] propose an inversion-based method called Action-Disentangled Identifier to learn action identifiers. Instead of a single identifier, the authors use one identifier per cross-attention layer and mask gradients unrelated to the action identifiers. Additionally, the authors present an action dataset called ActionBench on which similar models can be evaluated. In InteractDiffusion [35], rather than directly modifying the underlying model a control model is proposed which focuses on HOI (human object interaction) tasks.

More generally, ReVersion [36] performs relation inversion on generalized relationships that take the form of "subject 1 <relation> subject 2". The authors utilize the "preposition prior" which describes how prepositions typically describe relationships. Since parts of speech tend to be clustered in text-embedding space, the authors steer their relationship token towards the preposition cluster.

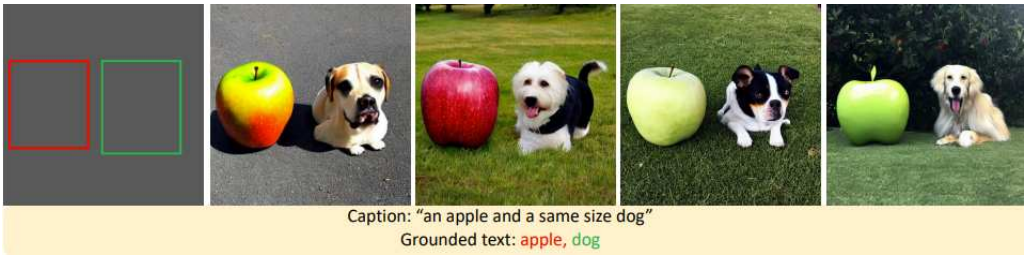

Figure 2: Examples of images generated by GLIGEN. Figure sourced from GLIGEN paper[37].

3.4 Layouts

Custom layouts are a common use case in controllable diffusion models. ControlNet [9] adds spatial controls on top of existing diffusion models by injecting zero-convolutions of a trainable copy into the frozen model. Various controls can be used, including edges, depth, segmentation, and pose. Similarly, GLIGEN [37] freezes model weights and injects grounding information (text prompts and bounding boxes) into new trainable layers via a gate mechanism. In both models, freezing weights preserves the vast knowledge obtained during the training of the base models. Afterwards, the same authors [38] implement a personalized version of this model that also disentangles object identity with other features such as location, which was a problem in prior models such as DreamBooth [7].

4 Open Problems

In this section, we will discuss existing problems in the field of controllable diffusion models, as well as what techniques have been proposed to address them.

4.1 Loss of Diversity and Catastrophic Forgetting

Due to the inherently low sample size for the custom concept task, models are at risk of overfitting, leading to loss of image diversity and forgetting of other concepts. Custom Diffusion [6] addresses this problem by recognizing that forgotten concepts are often related to the reference images. The authors use a regularization dataset of similar images (as defined by caption CLIP distance) during fine-tuning time, obtained either from the LAION-5B [39] dataset or generated using similar text prompts. Perfusion [40] utilizes a novel locking mechanism to restrict the influence of a new concept to cross-attention keys belonging to its category. SVDiff [41] avoid overfitting by compressing the parameter space, training only the singular values of the weight matrices.

4.2 Compositionality

A common problem with base models such as Dreambooth[7], Textual Inversion[8], and Custom Diffusion[6] are that they struggle with compositionality. Generated images may not respect spatial relationships ("A is on the left of B"), miscount objects, mix up colors between objects, or otherwise confuse subject attributes. This is particularly an issue in personalizable models and even more so when multiple personalized concepts are present. Figure 3 demonstrates how attributes between subjects can be confused by models like DreamBooth. While the prompt specifies that there should be both a cat and dog, only a dog is generated. Additionally, note that the attributes of the cat and dog are confused and applied to both subjects, resulting in inconstent identity of the uniquely identified "sks dog". While layout-based methods such as GLIGEN[37] can help to alleviate the spatial aspect of this problem, they require extra input and do not solve all related issues. Attend-and-excite [42] introduces Generative Semantic Nursing which more correctly bind attributes to their corresponding subjects by "exciting" the necessary activations, ensuring that all tokens are properly included and and result in greater textual alignment.

Attention Refocusing [22] also corrects attention maps, resulting in greater textual alignment with in a model-agnostic, training-free way. GPT-4 is used to generate prompt-compliant layouts, and attention of tokens are directed to focus on their corresponding bounding boxes and therefore more correctly reflect the prompt. Prompt Aligned Personalization (PALP) [43] more specifically addresses

Figure 3: Images generated by DreamBooth [7] with the prompt "cat next to an sks dog". Note how the number of subjects is incorrect on the left, and how the features of the cat and dog are mixed, with the dog dominating.

this issue for custom subject models like Dreambooth by retraining the model for a single prompt to ensure prompt alignment using score sampling. While prompt alignment is greatly improved, training is required for each custom subject and each new prompt and is therefore computation intensive.

5 Conclusion

In the few years since the introduction of DDPMs [1], the popularity of diffusion models for image generation has exploded. In this paper, we have provided an overview of underlying theory, applications, subproblems, and major developments in the field of personalizable text-to-image diffusion models. Furthermore, we address the present issues of image diversity and compositionality and recent papers addressing them. This survey aims to provide the reader with an understanding of the rapidly progressing field of personalizable T2I models and inform them of potential future areas of work.

References

[1] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

[2] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021.

[3] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[4] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T. Freeman, Michael Rubinstein, Yuanzhen Li, and Dilip Krishnan. Muse: Text-to-image generation via masked generative transformers, 2023.

[5] Pu Cao, Feng Zhou, Qing Song, and Lu Yang. Controllable generation with text-to-image diffusion models: A survey, 2024.

[6] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion, 2023.

[7] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation, 2023.

[8] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022.

[9] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

[10] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models, 2023.

[11] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications, 2024.

[12] Ziyi Chang, George Alex Koulieris, and Hubert P. H. Shum. On the design fundamentals of diffusion models: A survey, 2023.

[13] Hanqun Cao, Cheng Tan, Zhangyang Gao, Yilun Xu, Guangyong Chen, Pheng-Ann Heng, and Stan Z. Li. A survey on generative diffusion model, 2023.

[14] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10850–10869, September 2023.

[15] Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, and Yu-Gang Jiang. A survey on video diffusion models, 2023.

[16] Zhengbang Zhu, Hanye Zhao, Haoran He, Yichao Zhong, Shenyu Zhang, Haoquan Guo, Tingting Chen, and Weinan Zhang. Diffusion models for reinforcement learning: A survey, 2024.

[17] Yi Huang, Jiancheng Huang, Yifan Liu, Mingfu Yan, Jiaxi Lv, Jianzhuang Liu, Wei Xiong, He Zhang, Shifeng Chen, and Liangliang Cao. Diffusion model-based image editing: A survey, 2024.

[18] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2021.

[19] Calvin Luo. Understanding diffusion models: A unified perspective, 2022.

[20] Yutong He, Ruslan Salakhutdinov, and J. Zico Kolter. Localized text-to-image generation for free via cross attention control, 2023.

[21] Peiang Zhao, Han Li, Ruiyang Jin, and S. Kevin Zhou. Loco: Locally constrained training-free layout-to-image synthesis, 2024.

[22] Quynh Phung, Songwei Ge, and Jia-Bin Huang. Grounded text-to-image synthesis with attention refocusing, 2023.

[23] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

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

[26] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions, 2023.

[27] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[28] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control, 2022.

[29] Dan Song, Xuanpu Zhang, Juan Zhou, Weizhi Nie, Ruofeng Tong, Mohan Kankanhalli, and An-An Liu. Image-based virtual try-on: A survey, 2023.

[30] Yisol Choi, Sangkyung Kwak, Kyungmin Lee, Hyungwon Choi, and Jinwoo Shin. Improving diffusion models for virtual try-on, 2024.

[31] Li Chen, Mengyi Zhao, Yiheng Liu, Mingxu Ding, Yangyang Song, Shizun Wang, Xu Wang, Hao Yang, Jing Liu, Kang Du, and Min Zheng. Photoverse: Tuning-free image customization with text-to-image diffusion models, 2023.

[32] Zhuowei Chen, Shancheng Fang, Wei Liu, Qian He, Mengqi Huang, Yongdong Zhang, and Zhendong Mao. Dreamidentity: Improved editability for efficient face-identity preserved image generation, 2023.

[33] Dani Valevski, Danny Wasserman, Yossi Matias, and Yaniv Leviathan. Face0: Instantaneously conditioning a text-to-image model on a face, 2023.

[34] Siteng Huang, Biao Gong, Yutong Feng, Xi Chen, Yuqian Fu, Yu Liu, and Donglin Wang. Learning disentangled identifiers for action-customized text-to-image generation, 2024.

[35] Jiun Tian Hoe, Xudong Jiang, Chee Seng Chan, Yap-Peng Tan, and Weipeng Hu. Interactdiffusion: Interaction control in text-to-image diffusion models, 2024.

[36] Ziqi Huang, Tianxing Wu, Yuming Jiang, Kelvin C. K. Chan, and Ziwei Liu. Reversion: Diffusion-based relation inversion from images, 2023.

[37] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation, 2023.

[38] Yuheng Li, Haotian Liu, Yangming Wen, and Yong Jae Lee. Generate anything anywhere in any scene, 2023.

[39] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022.

[40] Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. Key-locked rank one editing for text-to-image personalization, 2023.

[41] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning, 2023.

[42] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models, 2023.

[43] Moab Arar, Andrey Voynov, Amir Hertz, Omri Avrahami, Shlomi Fruchter, Yael Pritch, Daniel Cohen-Or, and Ariel Shamir. Palp: Prompt aligned personalization of text-to-image models, 2024.